Vrije Universiteit Amsterdam


Air France-KLM


Master Thesis

# Forecasting upsell in the airline industry

**Author**: Caroline Oostveen (2675198)

*Daily KLM supervisors:* Tijs Klaver and Pieter Oorsprong
*First VU supervisor:* Joost Berkhout
*Second VU supervisor:* Rikkert Hindriks

*A thesis submitted in fulfillment of the requirements for the*
*VU Master of Science degree in Business Analytics (Optimization of Business Processes)*

# Preface

This study is conducted as part of the Master's program of Business Analytics at Vrije Universiteit Amsterdam. It has been carried out for the Air France-KLM offer management department and in collaboration with the Air France-KLM decision support department. The two main objectives are to find the driving factors behind the upsell KPI and to predict it accurately.

I would like to express my gratitude to my organizational supervisors, Pieter Oorsprong and Tijs Klaver from KLM, for the opportunity to conduct this research project and for their excellent guidance. Their assistance has been invaluable in offering direction and insight throughout the study. Furthermore, I would like to thank all my decision support colleagues for their support. Additionally, I would like to thank my supervisor, Joost Berkhout from the Vrije Universiteit Amsterdam, for his continuous guidance and encouragement throughout this research. Finally, I want to thank Rikkert Hindrinks for serving as my second reader.

## Summary

This study, conducted for Air France-KLM, aims to identify the factors influencing the upsell KPI and to forecast the upsell KPI for Air France and KLM. To achieve this, first, an explanatory data analysis has been performed to find out which variables have an impact on the upsell KPI. Next, several forecasts have been made to predict the upsell KPI. Two types of forecasts have been performed, namely a forecast that predicts in the near future, using information from passengers that have already booked the flight, and one that predicts in the far future, when no tickets have been sold yet.

In the exploratory data analysis, numerous variables have been found to have an impact on the upsell KPI, namely revenue group, day of week, travel motive, corporate, age group, length of stay, frequent flyer level, booking method, subclass, farebase season, traffic type and carrier. Using the feature importance graphs from the forecasts, it has also been found that the upsell KPI of the previous week, the distance and the week number are important. It was determined that random forest and XGBoost can accurately predict the upsell KPI. In addition, it has been found that the forecast including passenger information performs more accurately than the forecast without passenger information up until 41 weeks before departure.

# Contents

# 1 Introduction

## 1.1 Background

KLM Royal Dutch Airlines was founded in 1919 and is one of the oldest airlines globally [1]. KLM and KLM Cityhopper form the heart of the KLM group and have a total of 34.1 million passengers and 621.000 ton of Cargo yearly. KLM is a partner in the SkyTeam Alliance and has Transavia and Martinair as subsidiaries. In 2004, Air France and KLM decided to merge, resulting in the Air France-KLM group. They have defined their purpose as follows: 'At the forefront of a more responsible European aviation, we unite people for the world of tomorrow' [2].

Recently, the commercial departments of Air France and KLM have been seeking to increase their profit margins, resulting in a significant emphasis on upsell. They have defined upsell as follows: all extra revenue generated on top of a basic ticket price. They have defined three upsell categories, the ABC's: ancillary, branded fare, and cabin upsell.

- Ancillary upsell is simply all upsell coming from ancillary products and services, such as checked luggage, paid seat selection and several other products like WiFi and meals.
- Branded fare upsell is equal to the revenue from branded fares that are not the light fare, which is the most simple fare option. KLM has the following fare options: light, standard and flex. The differences between these fares are the checked baggage you are allowed to bring without paying fees, whether you can select your seats for free, whether you can make a change in your flights, and whether you can cancel your flight and get a refund.
- Cabin upsell refers to the income generated from tickets sold outside the economy class. KLM has three cabins: economy, premium comfort and business. Air France has the same cabins, but they have one additional cabin, namely the 'la première' cabin. The type of cabin determines your seating location on the plane, the comfort level of your seat, and the quality of meals and services provided during the flight.

Air France and KLM have introduced the upsell KPI to monitor the performance of the three upsell categories. This KPI is formulated as follows:

$$\text{upsell KPI} = \frac{\text{revenue from ancillaries } + \text{ revenue from branded fares upsell } + \text{ revenues from cabin upsell}}{\text{all ticket revenue}}$$

In this study, we will research the upsell KPI itself in order to better understand its underlying driving factors and evaluate if we can forecast the upsell KPI accurately in order to improve the KPI and thereby drive additional margin.

## 1.2 Motivation

Upsell plays a significant role in the airline industry. In recent years, more and more airlines are putting an emphasis on improving their upsell to improve their profitability. Given that airlines face competition from the low prices of budget carriers, focusing on upsell becomes crucial for generating profit. Legacy airlines like Air France and KLM, as well as Lufthansa and British Airways, are tapping into this domain.

Research shows that, without ancillary revenues, airlines would be operating at a loss [3]. Furthermore, it has been found that there is a positive correlation between airlines with a high percentage of revenue from ancillary products and services and those with high operating profits [4]. Since ancillary revenues are a component of upsell, it is crucial for Air France and KLM to closely examine and optimize their upsell strategies to enhance profitability.

## 1.3 Problem statement

KLM and Air France use the KPI metric to evaluate their performance. However, to date, not much research has been done by these airlines on this KPI. KLM and Air France want to investigate this KPI to understand which variables affect it, with the goal of finding key drivers of upsell performance and developing strategies to optimize their upsell revenue. Moreover, they want to have an improved way of setting a goal for the upsell KPI for the coming period. Right now, this is done by looking at previous years. A data analysis and forecast could help answer these questions and make Air France and KLM more knowledgeable about the upsell KPI. Air France and KLM want to be able to forecast for both short-term periods, when ticket sales are already ongoing, and long-term periods, before the ticket sales for a flight open. For forecasting when ticket sales are already ongoing, we can use information from the passengers that have already booked the flight. Therefore, this forecast will be referred to as the forecast with passenger information. The forecast before ticket sales open will be referred to as the forecast without passenger information.

The research questions that will be investigated are as follows.

**RQ1:** What factors are important when predicting how much upsell a customer will be doing?

1. Which categorical variables have a significant difference between the categories with respect to the upsell KPI corresponding to the different categories?

**RQ2:** Which models are suitable to forecast the upsell KPI?

1. Which models are most suitable for forecasting without passenger information?
2. Which models are most suitable for forecasting with passenger information?
3. How much time in advance is it better to use the forecast with passenger information instead of the forecast without passenger information?

**RQ3:** Can the driving factors and forecast be used and interpreted to increase the upsell revenue?

## 1.4 Outline

In Section 2, we will go over recent developments in upsell, discuss different forecasting methods, and discuss related research. In Section 3, we will perform an exploratory data analysis as well as discuss the different data sets and the preparation of these data sets. In Section 4, we will discuss the different methods used for forecasting. In Section 5, we will go over the different results for the forecasts that have been done. In Section 6, we will conclude the research and present some discussion points for future research.

# 2 Literature review

In this section, we discuss the literature review. In Section 2.1, we will go over all recent trends in upsell. In Section 2.2, we will discuss time series forecasting and how it can be used in this research. In Section 2.3, we will discuss the options for forecasting the continuation of a curve. In Section 2.4, we will go over the research papers related to this research.

## 2.1 Ongoing trends in upsell

The last few years, airlines have focused more on ancillary revenues. Ancillary revenue is described by O'Connell and Warnock-Smith as 'income beyond the sale of tickets that is generated by direct sales to passengers, or indirectly as a part of the travel experience' [5]. At the 2014 IATA World Passenger Symposium, senior IATA economists presented research showing a positive correlation between airlines with high operating profits as a percentage of revenue and airlines that had a high percentage of ancillary revenue [4]. Therefore, it would be good for airlines to focus more on upselling ancillary products.

The process of unbundling goes hand in hand with the increase in ancillary revenues [3]. Unbundling is the splitting of services that were once all part of a ticket, but now have to be booked separately. For example, hand baggage used to be part of an economy ticket, but now many airlines have changed it so that passengers have to pay extra for bringing hand baggage on board. Research found that passengers are less price sensitive when it comes to ancillary prices than when it comes to the ticket price itself [6]. This supports the need for airlines to focus more on ancillary products.

Ancillary revenues can be categorized in the following groups: a-la-carte ancillary revenues and third-party ancillaries. A-la-carte ancillaries are ancillaries that are sold by the airline directly to the passengers. Third-party ancillaries are ancillary products that are commission based and provided by third parties, such as car rental and travel insurance companies [7]. These third-party ancillaries are part of dynamic packaging. Dynamic packaging refers to booking complete travel

packages, such as flights, accommodation, transfers or tourist experiences [8].

## 2.2 Forecasting without passenger information

For the forecast without passenger information, there are two different types of forecasting models: time series models and machine learning models. Time series models, such as ARIMA, only take into account previous findings. For example, when dealing with sales data, they use the sales data of the previous years and make a forecast based on that, taking into account seasonality and trend. Machine learning models can consider numerous variables, including past observations, but also many more features. Machine learning models include linear regression, decision tree, random forest, XGBoost, and neural networks.

### 2.2.1 Time series models

The most commonly used time series model is the ARIMA model [9]. ARIMA stands for autoregressive integrated moving average. It assumes that there is a relation between the present value and the past values of the response variable. It also assumes that there is a relation between the current value of the series and past prediction errors. In the research performed by B. Pavlyshenko, it has been found that for sales prediction, machine learning models often perform better than time series methods due to the many patterns in the data [10]. The machine learning approach can better find these patterns in the data than the time series approach. Therefore, time series models are not used for this research.

### 2.2.2 Machine learning models

Linear regression [11] is the simplest machine learning model. It assumes that there is a linear relation between the response variable and one or more predictor variables, and uses this relation to make predictions.

The decision tree model [12] is also a relatively simple model. The main components are nodes and branches, hence why it is called a decision tree. Each node represents a decision that is made based

on features and each branch represents the outcome of such a decision. By doing so recursively, a set of rules is made that can be used to make predictions.

Random forest [13] is an ensemble learning technique that builds multiple decision trees during training. Their results are merged to improve accuracy and robustness, making them more accurate than decision trees. However, they are also less efficient and more time-consuming than the decision tree model.

XGBoost [14] is an implementation of gradient tree boosting, a machine learning technique that builds an ensemble of decision trees sequentially. The difference with random forest is that XGBoost builds trees sequentially, dependent on the previous trees, whilst the random forest algorithm does not take into account the other trees. Moreover, XGBoost is more efficient than random forest.

Neural networks are mathematical models, inspired on the functioning of biological neurons [15]. There are many different variations on neural networks. Remus and O'Connor [15] found that neural networks may work well for time series forecasting when working with monthly and quarterly time series, discontinuous series and for forecasts several periods into the future. Since we do not have solely monthly data, we have continuous data and we would also like to make predictions for the near future, we will not use neural networks.

## 2.3 Forecasting with passenger information

For the forecast that includes passenger information, we already know part of the curve for which we are trying to predict the end point, which is the upsell KPI in the week of departure. We can use these points as variables and use the machine learning models discussed in Section 2.2.2. Alternatively, we can use models specifically designed to forecast the continuation of a curve. These models only use the previous observations and not any other variable. Therefore, these models can serve as a benchmark to compare with models that include other variables, to see how much

these other variables contribute to the accuracy of the forecast. The most commonly used curve continuation models are listed below.

- The inflator algorithm [16] is an algorithm based on a multiplicative relation between the values of a curve. By first calculating the factor between two points, which are a certain number of time periods apart, and then smoothing this factor for stability, we can predict the endpoint of the curve.

- The additive algorithm [16] is similar to the inflator algorithm, but now it is assumed that there is an additive relation between the values of the curve. Similarly, first the difference between two values of the curve a certain number of time periods apart is calculated. After, this value is smoothed for stability and the forecast of the endpoint is calculated with the remaining value.

- The CuBaGe (curve base generator) model [17] is a model that will make a prediction based on other similar curves. There are historic full curves and partial curves for which we want to predict the continuation of the curve, or at least the last value of the curve. To forecast the continuation of the curve, the partial curves were compared with the historical full curves and the 10% most similar curves have been found and used.

## 2.4 Related research

Multiple researches have been done to find the willingness-to-pay of customers for economy class seat selection [18], ancillary services on long-haul flights [19] and dynamic packaging [8].

Yanfeng Zhou et al. [18] found that Chinese air consumers' willingness to pay is influenced by both intrinsic and extrinsic cues. Intrinsic cues that influence their willingness to pay are the length of the flight, seat comfort and convenience, and all of these have a positive impact on their willingness to pay for economy class seat selection. Extrinsic cues that influence the willingness to pay are payment and consumption situations. Consumers are sensitive to extra cost - if they have to pay a higher price to get a better seat, their willingness to pay will be lower. An example of a consumption

situation is traveling with friends and family. Customers want to sit together and therefore have a higher willingness to pay.

Paul Chiambaretto [19] found that long-haul passengers who travel mainly with a leisure travel motive are willing to pay more for ancillary services than those who travel with a business motive. Moreover, they found that the willingness-to-pay for a checked bag is a lot lower than the actual cost for checking a bag, while the willingness-to-pay for snacks and meals is similar to the actual cost.

Woon-Kyung Song and Hyun Cheol Lee [8] confirmed that Korean travelers have both the need and willingness to pay for dynamic packages offered by airlines. Convenience and economy are the main reasons for buying such dynamic packages. Airport transfers, foreign currency exchange, and travel insurance had both the highest need and the highest willingness to pay. Women, passengers below the age of 20 and the frequent flyers (who travel more than 10 times a year via air) have a significantly higher need for and willingness to pay a combination of ancillary products and services.

Moreover, researches have been performed to predicting the acceptance rate of premium airline seating [20] and upgrade offers [21].

Saravanan Thirumuruganathan et al. have researched predicting the upselling acceptance of premium airline seating [20]. A price elasticity model was implemented with two goals: (1) to identify the customers who are most likely to accept a seat upgrade offer, and (2) to determine the optimal price for these offers. This has been done based on the following variable categories: booking information (cabin class, booking class, family fare, flight date, booked date), customer demographics (age, gender, nationality), trip information (airport code, city, country of source and destination) and upgrade details (original and upgraded cabin, offer acceptance, offer price). This research can be used to target customers who are likely to accept an upgrade offer and to send them offers with the right upgrade price.

Noora Al Emadi et al. [21] have developed a model, called the PAX model, to predict passengers that purchase premium promotions. Their goal is to predict the passengers that are most likely to accept an upgrade offer. Their data consists of information about which customers were sent upgrade offers via e-mail and at what price they upgraded, if the customer decided to do so. They also use data about the nationality, gender and age of the passenger, the origin and destination of the flight, the flight date and ticket price. They found that both demographic information and price information are important for predicting the acceptance rate for upgrade offers.

To date, no other researches have been done concerning upsell. This literature review has given us several insights. From Section 2.1 on ongoing trends in upsell, we have learned how important upsell and specifically the ancillary upsell is for airlines. Additionally, by reviewing forecasting models in Section 2.2 and 2.3, we have identified which models are suitable for predicting the upsell KPI. The related research in Section 2.4 shows the results of various researches on upselling, each serving a different purpose than our study. Nevertheless, we can use several findings from these studies, such as which variables they found were important for predicting the acceptance of an upgrade offer.

# 3 Data analysis and data preparation

To start out data analysis, we will look more in-depth at the three categories of the upsell KPI. This will be done in Section 3.1. To start answering our first research question, exploratory data analysis will be performed to find out which variables have significant differences between their categories with respect to the upsell KPI. This will be done in Section 3.2. For question two of the research, we will make several forecasts. Before we can insert the data into the forecasting models, the data needs to be cleaned and transformed so that it is consistent, free of errors, and in a format suitable for accurate forecasting. How this will be done is explained in Section 3.3.

## 3.1 Upsell KPI

The upsell KPI can be split up into the three upsell categories which together make up the total upsell KPI. This is done by separately calculating the upsell KPI only for ancillary upsell, branded fare upsell, and cabin upsell. In Figure 1, we can see the different categories of upsell KPI, as well as the total upsell KPI, for each month in the year 2023. We see a drop in the cabin upsell KPI in the summer period. This is because in summer there is less corporate traffic. Corporate passengers make up a large part of the passengers who book a higher cabin, and thus the KPI will be lower when there are fewer corporate passengers. In the same summer period, we see an increase in the branded fare upsell KPI. This is because there are more leisure passengers in summer who want to bring more baggage on board, compared to the corporate passengers, who usually bring only hand luggage. We also see that the ancillary upsell KPI is more or less the same for the whole year. We see that the ancillary upsell KPI is between approximately [redacted]% and [redacted]%, the branded fare upsell KPI is between [redacted]% and [redacted]%, and the cabin upsell KPI is between [redacted]% and [redacted]%.

Figure 1: Upsell KPI per category per month for the year 2023

## 3.2 Exploratory data analysis

For both forecasts, the response variable is equal to the upsell KPI. Forecasts will be made for the upsell KPIs A, B, and C separately, as well as for the total upsell KPI. The exploratory data analysis will be performed only for the total upsell KPI.

### 3.2.1 Forecast without passenger information

For the forecast without passenger information, nine different datasets have been used. These datasets differ in their levels of aggregation. The data will be aggregated in two different ways: based on the flight level and on the time level. For the aggregation based on the flight level, three levels have been used: flight, subline, and complex group level. Flight means that all separate flights will be forecasted, a subline is all flights that have the same origin and destination, and a complex group consists of all flights to a certain region. For the time level, three levels were used as well: day, week, and month. In order to aggregate the data, the revenues are summed up over the time level, flight level, or both. For example, for the weekly subline level, the ancillary revenue, branded fare revenue, cabin revenue, and no upsell revenue are summed up over all days in a week and over all flights in a certain subline.

There are twelve variables that will be used in these datasets, but some of them cannot be used in all datasets. For example, for the weekly and monthly datasets, a variable based on a day (such as the day of the week) cannot be used. The nine datasets with their corresponding variables can be found in Table 1.

| | Flight | | | Subline | | | Complex | | |
|---|---|---|---|---|---|---|---|---|---|
| Variables | Daily | Weekly | Monthly | Daily | Weekly | Monthly | Daily | Weekly | Monthly |
| Revenue group | X | X | X | X | X | X | X | X | X |
| Complex group | | | | | | | X | X | X |
| Origin | X | X | X | X | X | X | | | |
| Destination | X | X | X | X | X | X | | | |
| Carrier | X | X | X | X | X | X | X | X | X |
| Aircraft owner | X | X | X | | | | | | |
| Scheduled arrival time | X | X | X | | | | | | |
| Scheduled departure time | X | X | X | | | | | | |
| Day of week | X | | | X | | | X | | |
| Day of year | X | | | X | | | X | | |
| Month | X | X | X | X | X | X | X | X | X |
| Year | X | X | X | X | X | X | X | X | X |

Table 1: Overview of explanatory variables included in the nine datasets for the forecast without passenger information

For the categorical variables revenue group and day of week, exploratory data analysis has been performed. We have investigated whether there are significant differences between these variables. For each variable, we first checked whether the distributions are approximately normally distributed to find out which test is appropriate to use. For normally distributed data, the one-way ANOVA test will be used to test whether there is a significant difference between the categories [22]. This test is the most suitable since it is used for a comparison of more than two unmatched groups and the response variable is numerical. For data that is not normally distributed, the Kruskal-Wallis test will be used [23]. For both examined variables, it has been found that there are significant differences between the variables. For the complete analysis that has been done, see Section A.1 in the Appendix. Next, a description of each variable will be given.

**Revenue group** Revenue groups are the different regions to which KLM and Air France fly. There are four revenue groups: Intercontinental, North Atlantic JV, medium haul and point to point. The intercontinental revenue group consists of all flights to and from Africa, Asia, Central and South America, the Middle East, the Gulf, the Indian Subcontinent, and the Caribbean and Indian Ocean. North Atlantic JV consists of all flights to and from Canada, Mexico, and the United States. Medium haul consists of all flights within Europe, and point to point includes all flights within France. All KLM flights fly from or to Amsterdam Schiphol Airport, and all Air France flights fly from or to Paris Charles de Gaulle Airport. These airports are called hubs.

**Day of week** The day of week variable indicates on which day the flight will depart. This is a categorical variable with the values Monday, Tuesday, Wednesday, Thursday, Friday, Saturday, Sunday.

**Origin and destination** The origin and destination variables are categorical variables. The categories are the airport codes of the departing and arriving airports of the corresponding flight.

**Scheduled departure time and scheduled arrival time** The scheduled departure time variable indicates the time that the corresponding flight is scheduled to depart and the scheduled arrival time indicates the time that the corresponding flight is scheduled to arrive. These are cyclical variables. The times are recorded in five-minute intervals.

**Day of year** The day of year variable is a numerical variable that indicates the day of the year that the corresponding flight is scheduled to depart. The variable ranges from 1 to 366. This variable is also cyclical, since the last day of one year and the first day of the next year are only one day apart (and not 364 days apart as suggested by the numbers 1 and 365).

**Month** The month variable is a numerical variable that indicates the month of the year that the flight departs in. It consists of the values 1 to 12. This variable is a cyclical variable as well.

**Year** The year variable is a numerical variable that states the year that the flight departs in.

**Aircraft owner** The aircraft owner variable states whether Air France-KLM is the owner of the aircraft and is therefore a binary variable. KLM and Air France sell some tickets for flights where passengers are transported by another operating carrier, for example flight AF5092 from Paris to Incheon International Airport, South Korea, which is operated by Korean Airlines. This is important, because only economy tickets can be sold and thus, the upsell KPI will be lower.

**Carrier** The carrier variable indicates the airline operating the flight. It is a categorical variable with the values Air France and KLM.

**Complex group** The complex group variable indicates which complex group the flight is part of. The complex group describes which region in the world the flight is flying to or from. There are 28 different complex groups, of which ten in Europe, seven in North-America, six in Asia, three in Africa and two in South-America.

### 3.2.2 Forecast with passenger information

For the forecast with passenger information, only one dataset will be used. This will be the weekly complex group dataset, as will be discussed in Section 5.1. Explanatory data analysis will be performed on the travel motive, booking method, traffic type, corporate, frequent flyer level, length of stay, subclass, age group and farebase season variables. The procedure of Section 3.2.1 will be used for this. It has been found that all of these variables have significant differences between their categories. See Section A.1 for the complete data analysis. We will give one example of this data analysis, which is for the booking method variable. The booking method variable provides information on how tickets were booked. It is a categorical variable with five different categories: direct offline, direct online, indirect offline, indirect online, and unknown. Direct online refers to passengers purchasing tickets, cabin upgrades, or ancillary services directly from the KLM website. Direct offline indicates that passengers purchased tickets at a KLM desk located at the airport.

Indirect offline refers to passengers buying tickets from an agent or travel agency. Lastly, indirect online denotes that tickets were purchased from an online travel agency. The unknown category states that it is not known how the passengers bought the tickets. A box plot has been made for the booking methods and their corresponding upsell KPI. This plot can be seen in Figure 2, where we see that there are quite some differences in the upsell KPI for the different categories. We especially see a large difference between the indirect online category and the other categories.

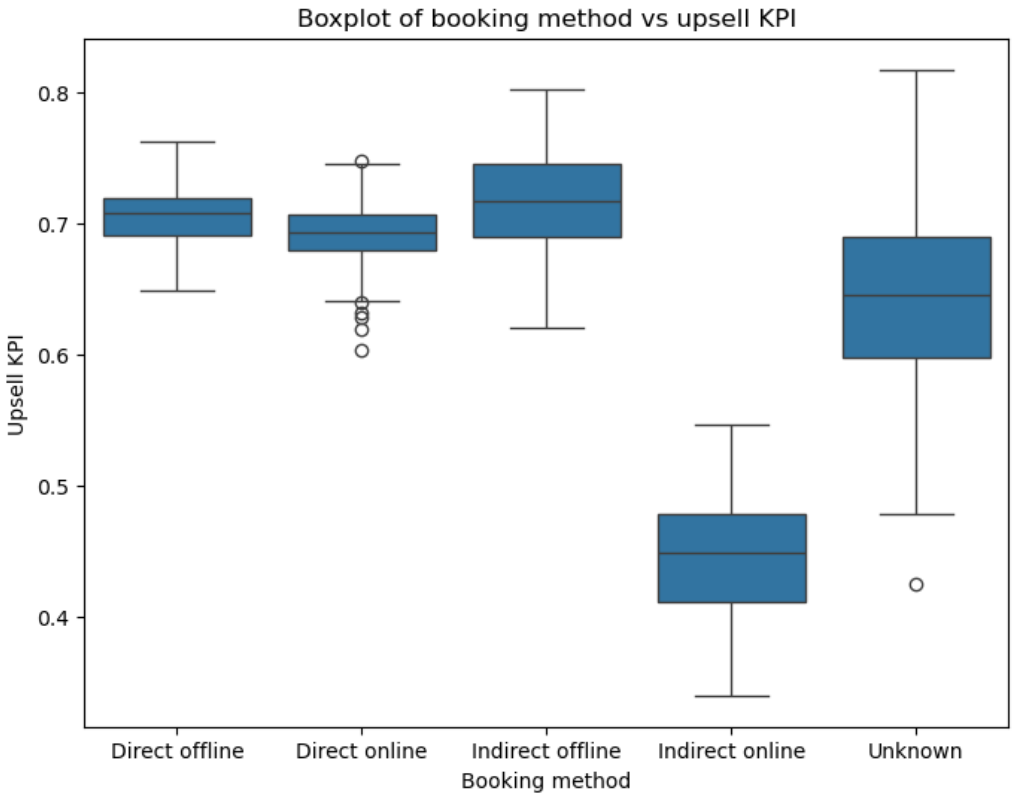


Figure 2: Box plot for booking method variable

First, we check for normality. For this purpose, histograms and Q-Q plots have been made. These can be found in figures 3 and 4. We can see that the data is indeed normally distributed. Next, a one-way ANOVA has been performed to check for a significant difference between the different booking methods. A p-value below 0.05 has been found, and thus it is concluded that there are

significant differences.

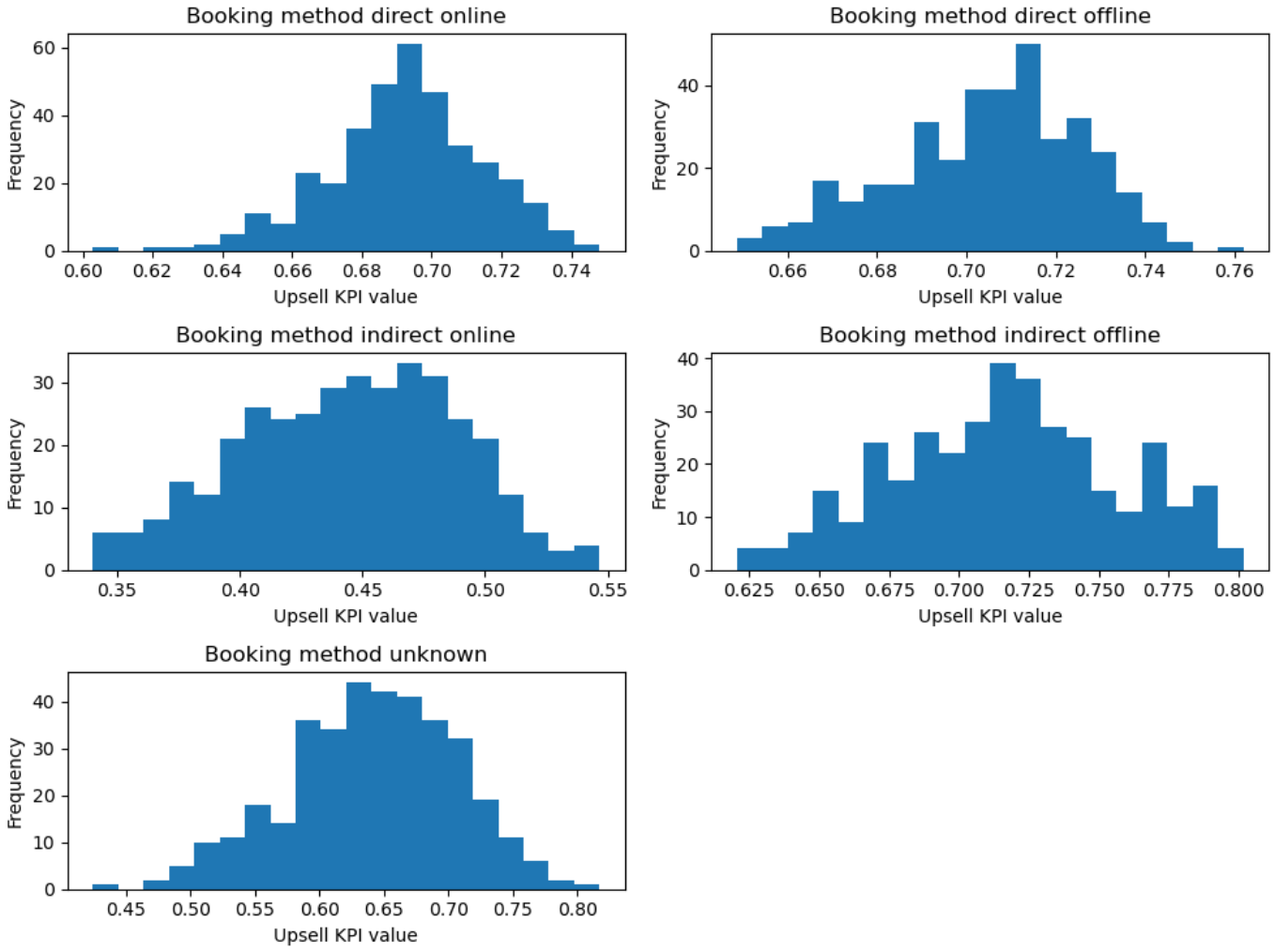


Figure 3: Histograms for booking method variable

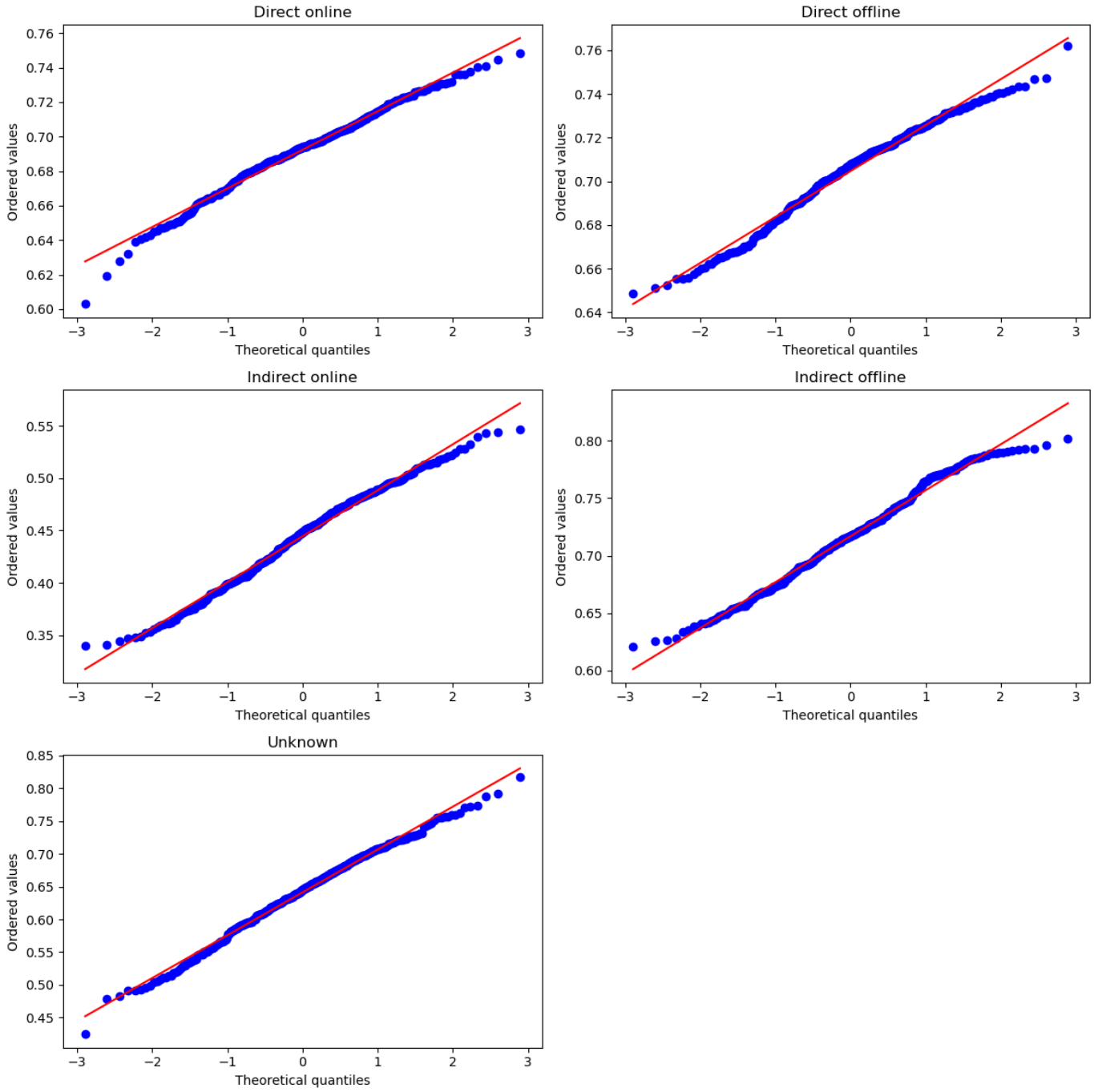


Figure 4: Q-Q plots for booking method variable

Next, a description of all other explanatory variables will be given.

**Travel motive** Travel motive is a categorical variable indicating the travel motive. There exist four travel motives: one way, business, leisure, and rest. There are certain rules attached to these travel motives. On the basis of these rules, the travel motive will be established. These rules differ for long- and medium-haul flights. For example, if there is no weekend stay and the duration of the stay is 2 days or less, for medium haul the travel motive will be set to business.

**Traffic type** The traffic type variable indicates what kind of connection a passenger has, if any. It is a categorical variable. The local category means that a passenger does not have a connecting flight, and the unknown category means that it is not known whether a passenger has a connecting flight. The other categories indicate the type of connection a passenger has: SH-SH, SH-MH, MH-MH, SH-LH, MH-LH or LH-LH. SH stands for short haul and they are all flights within France. MH stands for medium haul and consists of all flights within Europe. LH stands for long haul and consists of all flights outside of Europe. Thus, a SH-LH passenger is a passenger that first has a flight within France, and then a flight to another continent, for example the journey Nice Côte d'Azur Airport (NCE) to John F. Kennedy International Airport (JFK) with a layover at Paris-Charles de Gaulle Airport (CDG).

**Corporate** This variable is binary: a passenger is either a corporate passenger or not. Being a corporate passenger means that a corporate company has bought the ticket for this passenger.

**Frequent flyer levels** This variable is a categorical variable that states the frequent flyer level of the passenger if the passenger is in the frequent flyer program. KLM and Air France have a frequent flyer program for which passengers can enroll. By earning experience points they can go up in level. Passengers start at the explorer level and can then move up to the silver, gold and platinum levels accordingly. Each level brings with it its privileges and more earned miles per euro spent.

**Length of stay** Length of stay tells us how long the passenger will stay at their destination; how many days there are between their outbound and inbound flights. This is divided into the categories zero days, one day, two days, three days, four days, five days, six days, seven days or more, and unknown. The unknown category consists mainly of passengers who have booked only a one-way ticket and some of whom do not have enough data available to calculate the length of stay.

**Subclass** The subclass variable consists of 27 categories: the 26 letters of the alphabet representing a booking subclass and the unknown subclass category. Each ticket is booked in a certain subclass, which tells us its price category and cabin.

**Age group** The age group variable tells us something about the age of the passenger. This is a categorical variable that consists of three categories: infant, child, and adult.

**Farebase season** The farebase season variable consists of 8 categories: none, low, shoulder, high, peak, holiday, basic low low and unknown. It stand for which season the farebase is in. The farebase is a code which decides how high a ticket can be priced and the season states if it can be priced high or low depending on for example holidays.

**Upsell KPI** The upsell KPI of all previous weeks will be used as a variable to predict the upsell KPI of week $x$. Depending on the forecast the upsell KPI A, B, C or total will be used for this.

**Week** The week variable is a numerical variable ranging from 1 to 52. It indicates the week that the flight departs in. It is considered a cyclical variable since the week numbers return each year.

**Cabin** The cabin variable indicates the number of people that have booked a seat in each cabin. Four cabins exist: economy, premium comfort, business and la première. La première is only a cabin on Air France flights.

**Loadfactor and loadfactor per cabin** The loadfactor is equal to the percentage of seats that is booked, and the loadfactor per cabin is equal to the percentage of seats in each cabin that is booked. These variables say something about how many more passengers we can expect to book the flight. Important to note here is that sometimes, flights will be overbooked, and thus the percentage can be higher than 100%.

**Distance** The distance variable is the distance between the origin and the destination. If the distance is needed between a hub and a complex group, the average distance for all flights between the hub and the complex group in a year is taken.

In addition to these variables, the variables from Table 3.2.1 corresponding to the weekly complex group aggregation will also be used. These have been explained in Section 3.2.1.

## 3.3 Data preparation

In this section, we will prepare the data so that it is ready to be inserted in the forecasting models. In Section 3.3.1, we will explain how we will clean the data. In Section 3.3.2, we will explain how the data will be transformed. In Section 3.3.3, we will explain how the data is divided into training, testing, and validation sets.

### 3.3.1 Data cleaning

To clean the data, first of all, multi-leg flights have been removed. Multi-leg flights are flights that fly in a circle, for example AMS - CUR - AUA - AMS. This is the flight that departs from Amsterdam, stops in Curaçao, continues to Aruba, and then flies back to Amsterdam. In Curaçao, passengers with a ticket AMSCUR get out of the plane, while passengers with a ticket CURAMS get on the plane. At Aruba, passengers with an AMSAUA ticket get off the plane, while passengers with an AUAAMS ticket get on the plane. KLM does not sell tickets from CUR to AUA directly. These multi-leg flights are very hard to take into consideration, since the passengers on the plane have different tickets and it is difficult to divide the sales over the different flights. Therefore, it

has been decided to remove these flights from the dataset.

Furthermore, for several flights, data for arrival and departure times were missing. For flights with more than 25 missing times, we went on to search for the time and see if we could retrieve it using our other tools and list it manually. For flights with less than 25 missing times, these records were deleted.

There were also rows with negative values. It is not exactly known where these values come from; some can come from returned coupons, but mostly it is due to an error in the data. Therefore, all records with negative revenue values were deleted.

### 3.3.2 Transforming the data

The categorical variables revenue group, complex group and carrier were transformed with the use of one-hot encoding. In addition, the cyclical variables (scheduled arrival time, scheduled departure time, day of week, day of year, and month) were transformed into numerical variables such that the cyclic effect of these variables is taken into account. This has been done using the following formulae from [24]:

$$f_{cos}(x) = \cos\left(\frac{2\pi x}{\max(x)}\right) \tag{1}$$

$$f_{sin}(x) = \sin\left(\frac{2\pi x}{\max(x)}\right) \tag{2}$$

where $f_{cos}$ and $f_{sin}$ transform the data point $x$ into two different dimensions in cosine and sine. For example, for the variable 'day of week', the variables for day 3 are 0.434 for sine ($f_{sin}(3) = \sin(\frac{2\pi*3}{7})$) and -0.901 for cos ($f_{cos}(3) = \cos(\frac{2\pi*3}{7})$).

Moreover, the latitude and longitude of the origin and destination were found and included in the columns, and the names of the origin and destination were deleted. Lastly, all numeric variables

are normalized. This is done so that the scale of a variable is not taken into account. If one variable has bigger values than another, the model would focus more on the variable with the bigger values than the other variable. Therefore, all data is normalized with the min-max scaler so it is between the range 0 and 1. This is done by using the following formula from [25]:

$$f_{scaled}(x) = \frac{x - x_{min}}{x_{max} - x_{min}} \tag{3}$$

### 3.3.3 Train, test, and validation split

The goal will be to forecast values in the future when all values of the past are known, so an extrapolation. Therefore, training, validation, and test sets are also chosen such that the latest values are in the test set. The data consists of the years 2022, 2023 and the first four months of 2024. The four months in 2024 are used as a test set. For the data from 2022 and 2023, 25% is randomly sampled and makes up the validation set. The other 75% of this data is the training set.

# 4 Forecasting methods

In this section, we discuss the methods used to forecast the upsell KPIs A, B, C, and the total upsell KPI. In Section 4.1, the methods used for the forecast without passenger information are discussed. In Section 4.2, the methods used for the forecast with passenger information are discussed.

## 4.1 Forecast without passenger information

For the forecast without passenger information, four different models will be implemented: linear regression, decision tree, random forest, and XGBoost. In Section 2.2 of the literature review, we have elaborated on the reasons for selecting these methods.

### 4.1.1 Models

**Linear regression** Linear regression [11] is the simplest model considered. The linear regression model is represented by the following formula:

$$Y = \beta_0 + x_1\beta_1 + ... + x_p\beta_p + e$$

where

- $Y$ is the response variable, in our case the upsell KPI.
- $p$ is the number of predictor variables.
- $x_1, x_2, ..., x_p$ are the predictor variables; the variables as shown in Table 1.
- $\beta_0, ..., \beta_p$ are parameters.
- $e$ is the error term.

The formula represents a straight line with $\beta_0$ the intercept, and the coefficients $\beta_1, ..., \beta_p$ quantify the influence of the corresponding predictor variable on the response variable $Y$. The model will be trained on historical data to find the optimal values of the coefficients $\beta$, which will then be used to forecast the upsell KPI in the future.

**Decision tree** The decision tree model [12] is a more complex model. The main components are nodes and branches. There are three types of nodes:

- Root nodes that represent initial decisions that will result in the subdivision into other nodes.
- Internal nodes that represent possible choices leading to further subdivisions.
- Leaf nodes which represent the final result of the regression.

Branches connect nodes, forming paths from the root node to the leaf node. Each path represents a decision rule. Two methods used in the decision tree algorithm are splitting and stopping. Splitting is the process in which a parent node is split into multiple child nodes by making a decision based on an input variable. Stopping determines when a node should no longer be split, making it a leaf node. This is controlled by setting a maximum depth (the number of nodes from the root to a leaf) and a minimum number of samples required to split a node. If there are not enough samples or the maximum depth has been reached, the node will not split any further and this node will be the leaf node. In our forecasting problem, the decision tree will use historical data to learn the rules that lead to a certain upsell KPI. For example, the decision tree may split on a certain revenue group and carrier to split into a higher or lower upsell KPI. By following the decision paths from root node to leaf node, the model can predict the upsell KPI for new data.

**Random forest** Random forest [13] is an ensemble learning technique that builds multiple decision trees during training. Each tree in the forest is trained on a random subset of the training data and a random subset of the features. By training trees on different subsets of the training data, there will be diversity among the trees, which helps to reduce overfitting. During prediction, the

output of all trees is averaged to make the final prediction. In our forecasting problem, the random forest model will use historical data to learn patterns that influence the upsell KPI. By averaging the predictions from multiple decision trees, it is expected that a more accurate prediction will be made than that of an individual tree.

**XGBoost** XGBoost [14] is an implementation of gradient tree boosting, a machine learning technique that builds an ensemble of decision trees sequentially. Each tree is added to the ensemble to optimize a predefined loss function, gradually improving prediction accuracy with each addition. It uses regularization to avoid overfitting and takes advantage of parallel processing to speed up the training process. For our problem, decision trees will be built sequentially on our predictor variables, improving accuracy each step in order to forecast the upsell KPI accurately and efficiently.

These models will all be implemented four times, with the same explanatory variables but with different response variables. As explained in Section 1.1, the upsell KPI consists of three different parts that are added up to form the total upsell KPI. These three different parts, namely ancillaries, branded fares and cabin upsell, will all be forecasted separately. Next to that, the total upsell KPI will also be forecasted. Thus, there are four different response variables that will be forecasted: ancillary upsell KPI, branded fare upsell KPI, cabin upsell KPI and total upsell KPI. The explanatory variables can be found in Table 1 in Section 3.2.1.

### 4.1.2 Feature selection

The objective of feature selection is to find the subset of features with the lowest error. This provides a more robust generalization and makes the model more efficient [26]. We will use a feature selection method that is based on mutual information. Mutual information is a measure of the amount of information that one random variable has about another variable. It is equal to zero if and only if two variables are statistically independent. Mutual information gives a way to quantify how important a feature subset is with respect to the response variable. We have chosen mutual information for feature selection because of its efficiency and effectiveness. This method

allows us to evaluate a subset of features for every possible number of features without taking too long. The following steps are used for the process of feature selection using mutual information [26].

1. Mutual information scores are calculated between each individual explanatory variable and the response variable.
2. Features are sorted according to their mutual information scores in descending order.
3. For each possible number of features, ranging from 1 to the total number of variables in the dataset, one subset of features is selected. The subsets are chosen sequentially based on the order of the features. With this subset of features, cross-validation is performed to train the model and evaluate its performance. This performance evaluation is performed using the cross-validation score with the mean absolute error metric.
4. The feature subset with the minimal cross-validation score is chosen as the optimal feature subset.

In step 3, we make use of cross-validation. Cross-validation [27] is a data resampling method that is used to evaluate the ability of a model to generalize, and thus to prevent overfitting. A 5-fold cross-validation is used, which means that the training data set is divided into five disjoint subsets of the same size. Four of these sets are used for training and the fifth dataset is used for evaluation. This process is repeated five times, so that all folds are used once for evaluation. The mean average error over these five evaluation datasets is the performance measure.

Since the process of feature selection is time-consuming, first, the forecast will be done without feature selection. Then, feature selection will be performed for the model with the most accurate results.

### 4.1.3 Hyperparameter tuning

The performance of a machine learning model is dependent on its hyperparameters. Therefore, hyperparameter tuning [28] has been done in order to find the best hyperparameters. The random search technique, together with cross-validation, is used to tune the hyperparameters. For random

search, a set of possible values for each parameter is chosen, and the approach is to randomly sample from this set several times and find the best combination. This is done by training and evaluating with these parameters and using 5-fold cross-validation. The cross-validation score is used to compare the hyperparameter combinations.

For the XGBoost model, the following hyperparameters are tuned.

- The 'learning_rate' hyperparameter controls the step size at each iteration. Lower values will make the model take smaller steps, and thus will make the model go slower and more robust. The values 0.01, 0.1 and 0.2 are investigated.
- The 'max_depth' hyperparameter states the maximum length of a path in each tree. Deeper trees can capture more complex patterns, but can also lead to overfitting. The maximum depth levels 3, 5, and 7 are examined.
- The 'min_child_weight' hyperparameter is the minimum sum of instance weight needed in a child. This will prevent the model from creating too small leaves, which would make the model more complex. The levels 1, 3, and 5 are examined.
- The 'subsample' hyperparameter states the percentage of rows that is used for each tree construction. By lowering this value, the model will train on a smaller subset of the data and thus will be less likely to overfit. The percentages 50%, 70% and 100% are examined.
- The 'colsample_bytree' parameter is the percentage of features that is used for each tree construction. Again, lowering this value can make the model less prone to overfit by training on a subset of the variables. The percentages 50%, 70% and 100% are investigated.
- The 'gamma' parameter stands for the minimum loss reduction that is needed to split on a certain node. The values 0, 0.1, 0.5 and 1.0 are used.

For the random forest model, the following hyperparameters are tuned.

- The 'n_estimators' parameters is the number of trees in the forest. For this parameter, an interval of values is chosen, which is the interval [100, 300]. Whenever a value for this hyperparameter has to be chosen for the random search, the model will randomly choose an integer within this interval.
- The 'max_depth' hyperparameter is the maximum number of levels in each tree in the forest. When this hyperparameter is set too high, the model might overfit which leads to inaccurate results. Again, an interval is chosen for this parameter, which is the interval [10, 30].
- The 'min_samples_split' is the minimum number of data points that has to be used in a node before it can be split. Setting this value too low may result in overfitting, whereas setting it too high could lead to underfitting. The interval [1, 10] is chosen for this hyperparameter.
- The 'min_samples_leaf' hyperparameter is the minimum number of data points that have to be in a leaf node. Larger values of this parameter create more generalized trees by making sure that leaf nodes have enough data points. The small interval [1, 4] will be used for this parameter.
- The 'bootstrap' hyperparameter is a binary parameter that states whether we sample data points with or without replacement. The two values that are evaluated for this parameter are true and false.

Hyperparameter tuning is a time-consuming process. Therefore, it has been decided to only perform hyperparameter tuning on the model with the most accurate results.

## 4.2 Forecast with passenger information

### 4.2.1 Models

For this type of forecast, the data from Section 3.2 has been used. Since this forecast is more advanced and therefore takes more time to build, it has been decided to perform this forecast for only

one aggregation. Based on both the results of the flight forecast (to be discussed in Section 5.1) and the opinion of the analysts who want to use the forecast, it has been decided to continue with the weekly complex group aggregation. This has been chosen since analysts will not look at day - flight level at the upsell KPI, but at a higher level. The weekly complex level is a good trade-off between accurate results and usability.

Firstly, the same four models that were used in Section 4.1 will be used, namely linear regression, decision tree, random forest and XGBoost. The same four response variables will be used, namely the ancillary upsell KPI, branded fare upsell KPI, cabin upsell KPI and the total upsell KPI. The explanatory variables are the variables explained in Section 3.2. The same process for feature selection and hyperparameter tuning is used as for the forecast without passenger information, see Section 4.1.2 and 4.1.3, respectively.

Next to that, three other models were implemented: the inflator algorithm, the additive algorithm and the CuBaGe algorithm. These models only use past observations to make their predictions. They can be used as a benchmark; we can compare how much more accurate we can forecast when using both flight information and past observations compared to only using past observations. The past observations consist of the upsell KPI for the realised ticket sales for all weeks before departure, which are 53 weeks. Next, we will describe the three benchmark methods.

**Inflator algorithm** The inflator algorithm [16] is an algorithm based on a multiplicative relation between the values of a curve, in our case the curve of the upsell KPI values in all weeks before departure. When we want to predict a certain value $S_t^j$, which is the value of the upsell KPI $S$ at time $t + j$ predicted from time $t$, we do the following. First, we divide the current value of the upsell KPI $S_t$ by the value realized at time $t - j$, which is $S_t^{-j}$, and call this $\gamma_t^j$.

$$\gamma_t^j = \frac{S_t}{S_t^{-j}} \tag{4}$$

$\gamma_t^j$ is called the scale-up factor. This scale-up factor is then smoothed for stability:

$$\psi_t^j = (1-\alpha)\psi_{t-1}^j + \alpha\gamma_t^j \tag{5}$$

where $\psi_t^j$ is called the inflating factor and $\psi_0^j = \gamma_0^j$. .
Then, the value of the upsell KPI $S$ for period $t+j$ is estimated as follows:

$$S_t^j = \psi_t^j S_t \tag{6}$$

**Additive algorithm** The additive algorithm [16] is similar to the inflator algorithm, but now it is assumed that there is an additive relation between the value of the curve now and the value for the same period $x$ amount of time in advance. Again, we want to predict the value of the upsell KPI $S$ for time $t+j$ from time $t$. At the current time $t$, we know a certain part of the upsell KPI, namely $S_t$. The unknown part will be estimated as follows: from the value of the upsell KPI at time $t$, we subtract the known portion of the upsell KPI $j$ times earlier; so at $t-j$.

$$u_t^j = S_t - S_t^{-j} \tag{7}$$

where $u_t^j$ is the unknown part of the upsell KPI for period $t$ in period $t-j$. The value of $u_t^j$ is smoothed as follows:

$$F_t^j = (1 - \alpha)F_{t-1}^j + \alpha u_t^j \tag{8}$$

where $F_0^j = u_t^j$.

Lastly, the value of the upsell KPI for time $t + j$ is estimated as follows

$$S_t^j = F_t^j + S_t \tag{9}$$

**CuBaGe** The CuBaGe (curve base generator) model [17] is a model that will make a prediction based on other similar curves. There are historic full curves and partial curves for which we want to predict the continuation of the curve, or, in our case, only the last value of the curve, which is the final upsell KPI. To forecast the continuation of the curve, the partial curves were compared with all other curves, and the 10% most similar curves have been found and used. To compare partial curves with full curves, only the first part of the full curve, matching the length of the partial curve, will be used. The similarity is based on the distance between the curves, namely with the following formula:

$$\text{sim}(c, c') = \frac{1}{1 + \text{dist}(c, c')} \tag{10}$$

where $c$ and $c'$ are the two curves we are comparing and $\text{dist}(c, c')$ is the distance between these curves. The distance between curves is calculated as follows. For each time point, we calculate the difference between the values of the two curves at that point. These differences are squared, and we sum up the squared differences of all time points. Then we take the square root of this sum and

this value we use as the distance between the two curves.

For every curve $i$, a goodness score is also calculated, which decides how comparable the curve $i$ and the curve for which we are making a prediction, curve $c$, are with respect to all other curves. The goodness score is then used to determine the weight assigned to each point, influencing its impact on the forecast. This is done with the following formula:

$$\text{goodness}_i = \frac{\text{sim}(i, c)}{\sum_{\text{all}_j} \text{sim}(c, j)} \tag{11}$$

Then, to predict what will become the final point of curve $i$, thus the final upsell KPI, the following is done:

$$\hat{u_i} = \sum_{\text{10\% most similar curves } k} \text{goodness}_k * u_k \tag{12}$$

where $u_k$ is the last point of the complete curve $k$, which matches the point we are trying to predict for curve $i$ timewise.

# 5 Results

In this section, we will go over all the results. The results include prediction errors, box plots of prediction errors, confidence intervals, Wilcoxon tests results, and feature importance graphs. In Section 5.1, we will discuss the results of the forecast without passenger information. In Section 5.2, we show the results of the forecast including passenger information.

## 5.1 Forecast without passenger information

For the forecast without passenger information, first, only a forecast for the total upsell KPI has been made. The results of this forecast, for the four forecasting models linear regression, decision tree, random forest, XGBoost, and the nine datasets, can be found in Table 2. The error measure used is the mean absolute error, which indicates the average deviation of the forecasted percentage from the actual percentage.

| | Linear regression | Decision tree | Random forest | XGBoost |
|---|---|---|---|---|
| Daily flight | 9.6315% | 11.4297% | 8.4115% | 8.4529% |
| Weekly flight | 9.2003% | 8.9848% | 7.3118% | 7.2463% |
| Monthly flight | 6.6919% | 6.9809% | 5.7467% | 6.2629% |
| Daily subline | 8.5783% | 9.8569% | 7.4532% | 7.2265% |
| Weekly subline | 7.0899% | 6.6983% | 6.0174% | 5.8961% |
| Monthly subline | 6.6734% | 6.2762% | 5.6164% | 5.4791% |
| Daily complex group | 3.8077% | 4.4449% | 3.6949% | 3.7291% |
| Weekly complex group | 3.2761% | 3.4324% | 3.2326% | 3.1671% |
| Monthly complex group | 3.3396% | 4.0365% | 3.6700% | 3.7528% |

Table 2: Results for the forecast without passenger information with mean absolute error as error measure

We can see that there are quite big differences between the models for the same aggregations. The Wilcoxon test has been used to find whether the differences between the models are statistically significant [29]. This test is used since we have paired comparisons, given that the predictions are made on the same data. We tested the differences between all models and we found the following results.

- There is no significant difference between the linear regression model and the decision tree

model.

- The random forest model performs significantly better than both the linear regression model and the decision tree model.
- The XGBoost model performs significantly better than both the linear regression model and the decision tree model.
- There is no significant difference between the random forest model and the XGBoost model.

We can see that weekly complex group is the data aggregation that has the most accurate predictions. Therefore, it has been decided to only continue with this data aggregation. The following results will therefore only be shown for the weekly complex group. For this data aggregation, the forecast for upsell KPI A, upsell KPI B and upsell KPI C has also been performed. These separate predictions are added up to compare with forecasting the total upsell KPI at once. The results can be found in Table 3. We can see that there is only a small difference between forecasting the upsell KPI at once and forecasting the three upsell KPIs separately and then adding up these predictions.

| | Linear regression | Decision tree | Random forest | XGBoost |
|---|---|---|---|---|
| Upsell KPI A | 0.5817% | 0.6071% | 0.5539% | 0.5561% |
| Upsell KPI B | 4.5761% | 3.6944% | 3.5030% | 3.4421% |
| Upsell KPI C | 3.9324% | 3.0450% | 2.8030% | 2.7292% |
| Upsell KPI total | 3.2761% | 3.4324% | 3.2326% | 3.1671% |
| Upsell KPI added up | 3.2766% | 3.5438% | 3.2903% | 3.2385% |

Table 3: Results weekly complex group for forecasting upsell KPI A, B, C and the total upsell KPI

A Wilcoxon test was performed again to check whether there is a significant difference between predicting the total upsell KPI at once and adding up the predictions for the upsell KPIs A, B and C. It has been found that there is no significant difference between these two results. This means that we can use the predictions for the upsell KPIs A, B, and C and add these to find the total upsell KPI without loss of accuracy. Analysts can use this to look at the forecasts of A, B, and C separately and use this information to see which of the upsell categories is lagging behind or going well.

For the weekly complex total forecast, a box plot has been made including the mean absolute errors for the four forecasting models. This box plot can be found in Figure 5.

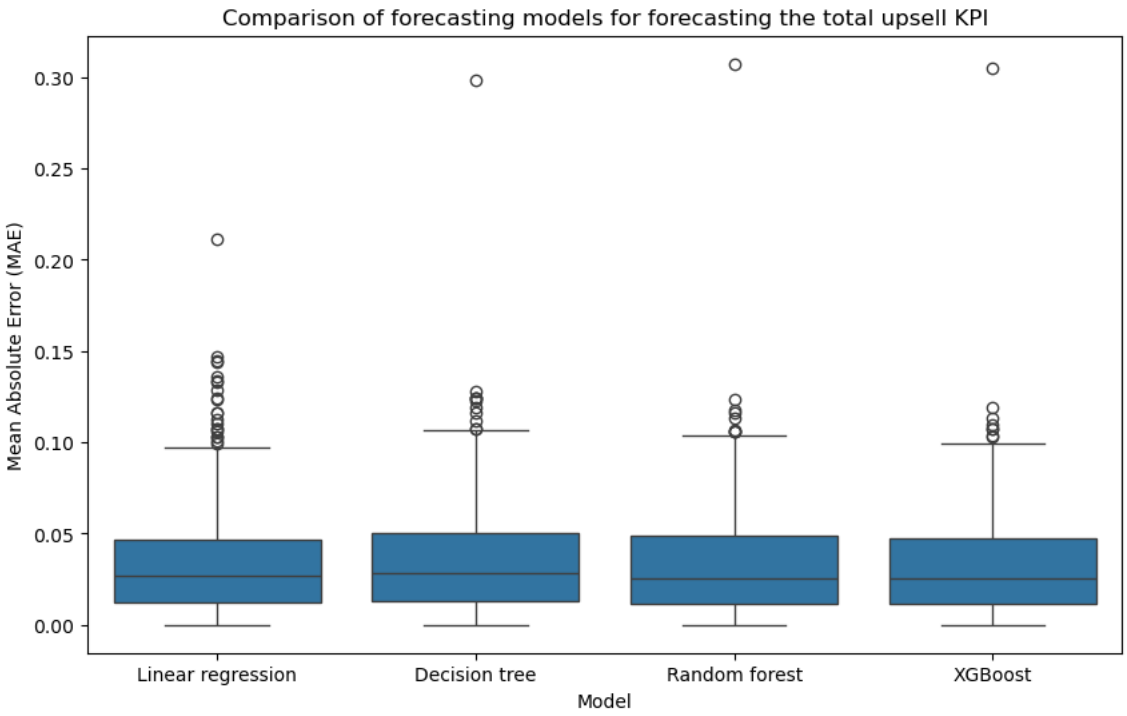


Figure 5: Boxplot for the results for the four forecasting models for forecasting the total upsell KPI

In addition, 95% confidence intervals have been constructed for each prediction. The spread of these intervals can be found in Table 4. For example, if the prediction for the total upsell KPI for the XGBoost model is 74%, then with 95% certainty, the actual value will be between 74.27% and 73.73%.

| | Linear regression | Decision tree | Random forest | XGBoost |
|---|---|---|---|---|
| Upsell KPI A | 0.1007% | 0.1072% | 0.0979% | 0.0968% |
| Upsell KPI B | 0.7183% | 0.6141% | 0.5894% | 0.5802% |
| Upsell KPI C | 0.6406% | 0.5541% | 0.5116% | 0.4890% |
| Upsell KPI total | 0.5610% | 0.5800% | 0.5557% | 0.5423% |
| Upsell KPI added up | 0.5611% | 0.5980% | 0.5590% | 0.5514% |

Table 4: Spread for the 95% intervals for the weekly complex group aggregation for the four forecasting models

The XGBoost model has the most accurate results. Therefore, feature selection and hyperparameter tuning has been performed for the XGBoost model. The results can be found in Table 5.

| | XGBoost |
|---|---|
| Upsell KPI A | 0.5686% |
| Upsell KPI B | 3.4824% |
| Upsell KPI C | 2.7217% |
| Upsell KPI total | 3.1718% |
| Upsell KPI added up | 3.2301% |

Table 5: Results for the XGBoost model after feature selection and hyperparameter tuning

It has been found that for all models, the most accurate feature subset is the complete set of features. We can see that some models improve due to the tuning of hyperparameters, but other models become less accurate. This means that, for the validation set, other hyperparameters than the default parameters are optimal. However, for the test set, these hyperparameters actually perform worse. This can happen, since the test set has never been seen before by the model.

Moreover, the feature importance graphs for the XGBoost model for upsell KPIs A, B, C and the total upsell KPI will be shown. They can be found in Figures 6, 7, 8 and 9. The feature importance graphs for the other models have been included in the Appendix sections A.2.1, A.2.2, A.2.3 and A.2.4.

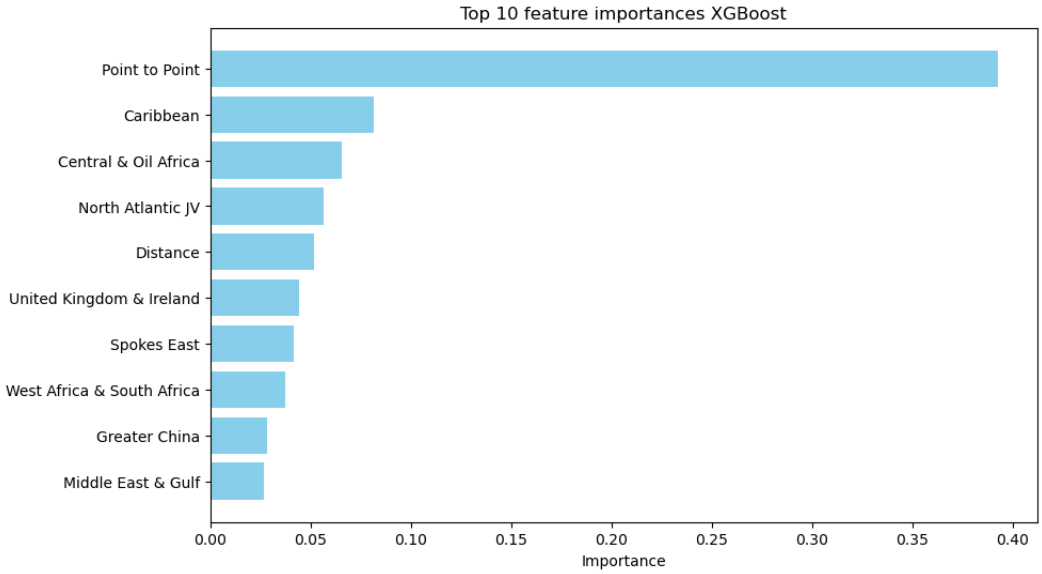


Figure 6: Feature importance for XGBoost for forecasting upsell KPI A

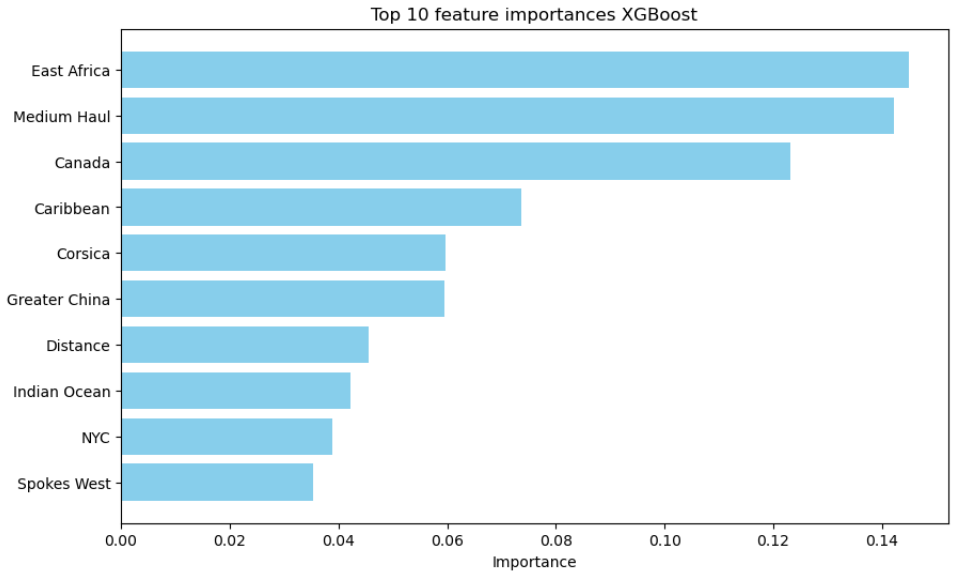


Figure 7: Feature importance for XGBoost for forecasting upsell KPI B

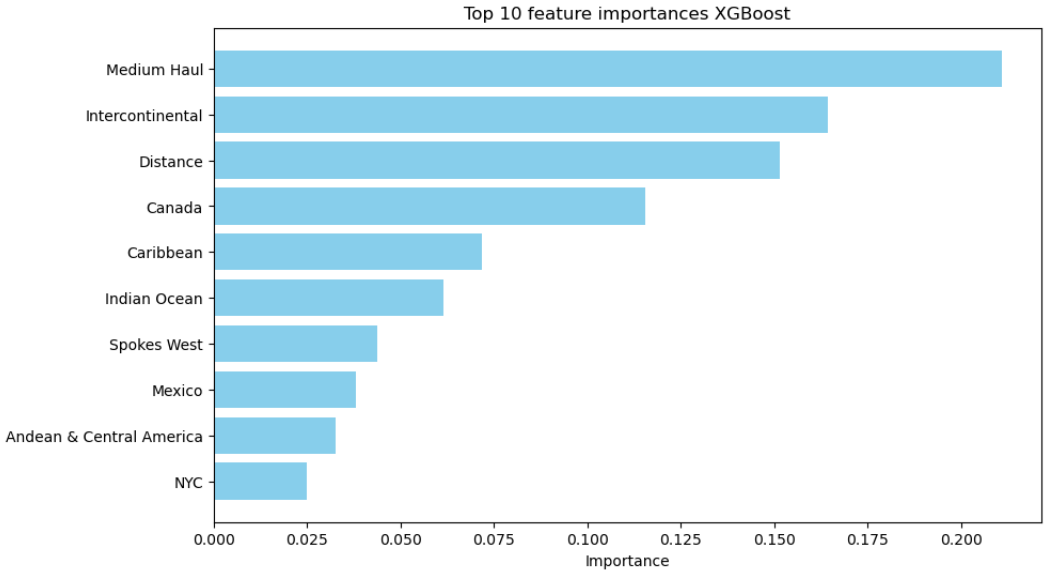


Figure 8: Feature importance for XGBoost for forecasting upsell KPI C

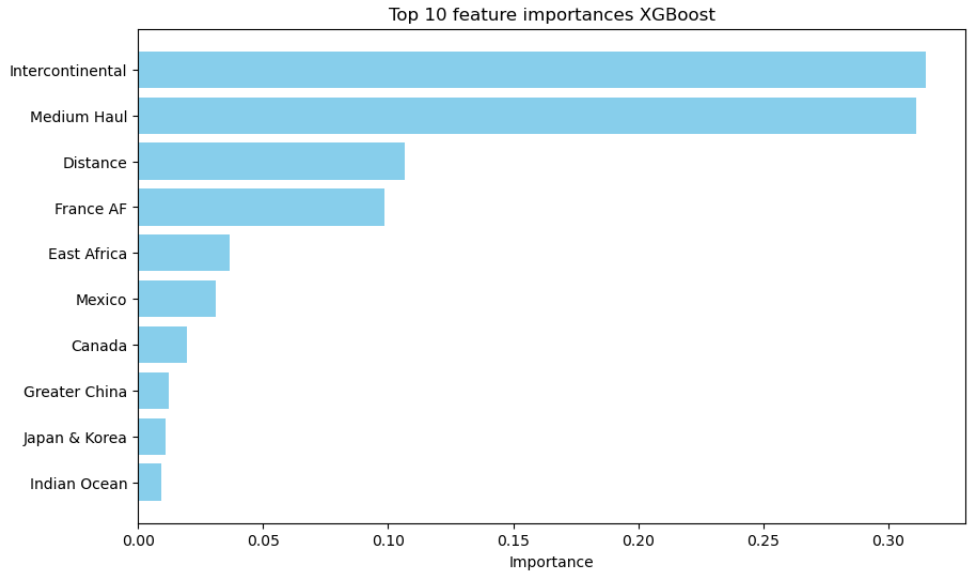


Figure 9: Feature importance for XGBoost for forecasting the total upsell KPI

We can see that the different models also find different features more important. For predicting the ancillary upsell KPI, the revenue group point to point is very important, namely the most important feature with an importance of almost 0.4. This feature is not in the top 10 of any other model.

## 5.2 Forecast including passenger information

For the forecast including passenger information, we expect to find better results than for the forecast without passenger information, at least for several weeks before departure. This forecast will be performed using the same methods as for the forecast without passenger information, but we will also use the three curve-based models that are discussed in Section 4.2. The forecast was first performed for all seven models for the total upsell KPI only. For the inflator and additive algorithms, we need at least double the number of weeks that we are predicting. Thus, these models can only predict up to 26 weeks before departure. The CuBaGe model needs at least two data points for the curve, and thus this model cannot predict 52 weeks in advance. The results can be found in Table 6.

| Weeks in advance | Linear regression | Random forest | Decision tree | XGBoost | Inflator | Additive | CuBaGe |
|---|---|---|---|---|---|---|---|
| 1 | 0.7577% | 0.5221% | 0.7370% | 0.5499% | 0.6036% | 0.6132% | 2.3940% |
| 2 | 1.2157% | 1.0299% | 1.3490% | 1.0457% | 1.0142% | 1.0275% | 2.5624% |
| 3 | 1.5583% | 1.4037% | 1.8273% | 1.3456% | 1.4176% | 1.4149% | 2.7282% |
| 4 | 1.8230% | 1.6122% | 2.1331% | 1.5123% | 1.6693% | 1.6782% | 2.8893% |
| 5 | 2.0420% | 1.7371% | 2.3230% | 1.7857% | 2.0385% | 2.0387% | 3.0491% |
| 6 | 2.2905% | 1.9067% | 2.4716% | 1.9985% | 2.3642% | 2.3188% | 3.1992% |
| 7 | 2.5892% | 2.0536% | 2.5563% | 2.0370% | 2.6052% | 2.5456% | 3.3444% |
| 8 | 2.7851% | 2.3221% | 3.0007% | 2.3405% | 2.8159% | 2.7489% | 3.4783% |
| 9 | 2.9081% | 2.4235% | 3.1332% | 2.3837% | 2.9935% | 2.9002% | 3.6082% |
| 10 | 2.9925% | 2.4983% | 3.4600% | 2.5043% | 3.1484% | 3.0607% | 3.7348% |
| 11 | 3.0551% | 2.5282% | 3.1725% | 2.5888% | 3.3253% | 3.2214% | 3.8711% |
| 12 | 3.1111% | 2.5395% | 3.5110% | 2.6171% | 3.4900% | 3.3834% | 3.9889% |
| 13 | 3.1748% | 2.5613% | 3.3629% | 2.7319% | 3.7064% | 3.6017% | 4.1092% |
| 14 | 3.2125% | 2.7038% | 3.5624% | 2.8107% | 3.9242% | 3.8106% | 4.2325% |
| 15 | 3.2173% | 2.7399% | 3.6719% | 2.9398% | 4.1863% | 4.0527% | 4.3559% |
| 16 | 3.2160% | 2.7578% | 3.5896% | 2.9418% | 4.5284% | 4.3711% | 4.4675% |
| 17 | 3.2225% | 2.7817% | 3.6217% | 2.7603% | 4.9178% | 4.6967% | 4.5904% |
| 18 | 3.2308% | 2.8573% | 3.5716% | 2.9217% | 5.7375% | 5.1168% | 4.7215% |
| 19 | 3.2410% | 2.8655% | 3.8227% | 2.9721% | 5.9080% | 5.4539% | 4.8636% |
| 20 | 3.2579% | 2.8819% | 3.8332% | 3.0619% | 6.5430% | 5.9124% | 4.9889% |
| 21 | 3.2549% | 2.9058% | 3.8586% | 2.9859% | 7.0937% | 6.5312% | 5.1131% |
| 22 | 3.2587% | 2.8674% | 3.7982% | 2.9363% | 8.2511% | 7.2098% | 5.2370% |
| 23 | 3.2692% | 2.8976% | 3.8253% | 3.0349% | 12.8986% | 8.0324% | 5.3727% |
| 24 | 3.2605% | 2.9609% | 4.0643% | 3.0852% | 15.4686% | 9.9215% | 5.4896% |
| 25 | 3.2868% | 2.9928% | 3.7694% | 3.0109% | 19.2080% | 13.0121% | 5.5791% |
| 26 | 3.3035% | 2.9324% | 3.7497% | 3.0727% | 51.2356% | 26.2683% | 5.6663% |
| 27 | 3.3493% | 2.9678% | 3.6601% | 2.8726% | | | 5.7795% |
| 28 | 3.3764% | 2.9361% | 3.8258% | 3.0124% | | | 5.8754% |
| 29 | 3.4226% | 2.8740% | 3.6182% | 3.1417% | | | 5.9811% |
| 30 | 3.4679% | 2.9215% | 3.6108% | 3.1473% | | | 6.0986% |
| 31 | 3.4961% | 2.9325% | 3.7038% | 3.0759% | | | 6.2300% |
| 32 | 3.5235% | 2.9064% | 3.8742% | 3.1964% | | | 6.3208% |
| 33 | 3.5623% | 3.0299% | 3.8171% | 3.1411% | | | 6.4469% |
| 34 | 3.5899% | 3.0003% | 4.0146% | 3.2272% | | | 6.5697% |
| 35 | 3.6303% | 3.0245% | 3.9060% | 3.2497% | | | 6.6790% |
| 36 | 3.6551% | 3.0825% | 4.0145% | 3.3163% | | | 6.7715% |
| 37 | 3.6931% | 3.1499% | 4.0464% | 3.4226% | | | 6.8551% |
| 38 | 3.7100% | 3.2366% | 4.0242% | 3.4208% | | | 7.0436% |
| 39 | 3.7273% | 3.2440% | 3.9121% | 3.3560% | | | 7.0488% |
| 40 | 3.7477% | 3.2900% | 3.9572% | 3.3259% | | | 7.1653% |
| 41 | 3.7581% | 3.4244% | 4.6335% | 3.4223% | | | 7.3427% |
| 42 | 3.7691% | 3.4251% | 4.4876% | 3.4535% | | | 7.4245% |
| 43 | 3.7916% | 3.4712% | 4.3525% | 3.5224% | | | 7.4864% |
| 44 | 3.8008% | 3.4982% | 4.6413% | 3.6081% | | | 7.5460% |
| 45 | 3.8233% | 3.5194% | 4.3750% | 3.4972% | | | 7.7863% |
| 46 | 3.8490% | 3.5546% | 4.1384% | 3.5396% | | | 7.7553% |
| 47 | 3.8568% | 3.5870% | 4.1522% | 3.5254% | | | 8.1260% |
| 48 | 3.8636% | 3.6559% | 4.5614% | 3.7262% | | | 8.2334% |
| 49 | 3.8737% | 3.8218% | 4.6185% | 3.6894% | | | 8.8944% |
| 50 | 3.8697% | 3.8626% | 4.8312% | 3.7313% | | | 8.6560% |
| 51 | 3.8997% | 3.9163% | 4.8538% | 4.0008% | | | 9.3229% |
| 52 | 3.9733% | 4.1312% | 4.7093% | 4.0475% | | | |

Table 6: Results for the forecast with passenger information with mean absolute error as error measure

We can see that the machine learning models are more accurate than the curve based models. For 20 weeks before departure, the most accurate machine learning model has a 2.11 percent more accurate prediction than the most accurate curve based model. Thus, the variables that are not used in the curve based model, which are all variables except for the past observations of the upsell KPI, improve the prediction by 2.11%. A boxplot has been made for the four machine learning models, which can be found in Figure 10.

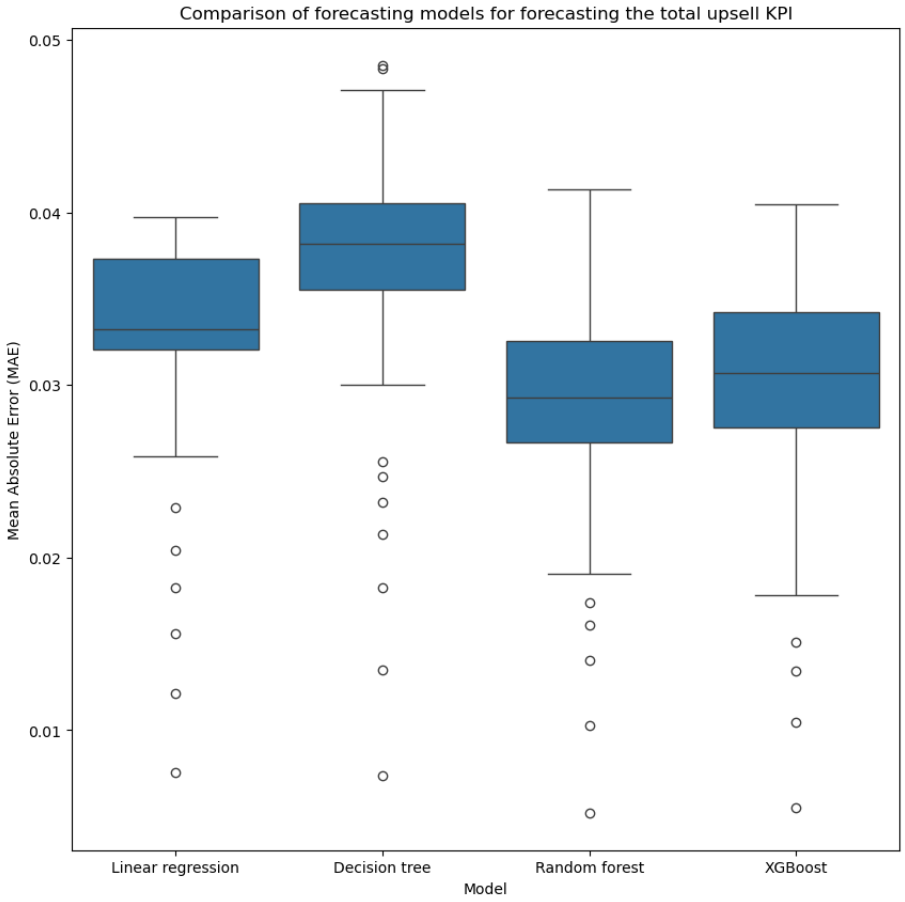


Figure 10: Boxplot for the results for the machine learning forecasting models for forecasting the total upsell KPI

Again, several Wilcoxon tests have been performed to check for significant differences between the models. The outcomes from the tests are depicted in Table 7. Interpretation of this table starts from the left: if a green color appears under a column for a specific row, it indicates that the model in the row is significantly better than the one in the column. A red colour means that the row model is significantly worse than the column model, and an orange colour means there is no significant difference between the models.

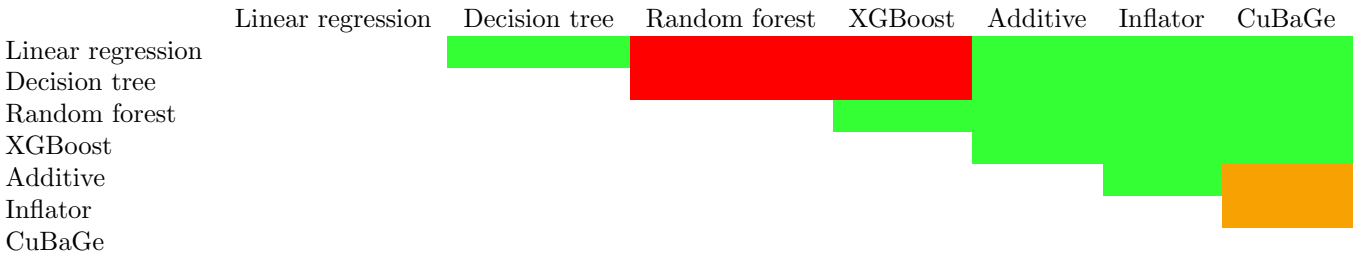


Table 7: Outcome Wilcoxon tests for significant differences between the models for the forecast including passenger information

It can be seen that the random forest model is significantly better than all other models. Therefore, we continue only with this model. We have made predictions for the upsell KPIs A, B and C separately, as well as the total upsell KPI. Moreover, for all these models, we have performed feature selection and hyperparameter tuning. The results of these advanced models can be found in Table 8.

| Weeks in advance | Upsell KPI A | Upsell KPI B | Upsell KPI C | Total upsell KPI | Upsell KPI added up |
|---|---|---|---|---|---|
| 1 | 0.3831% | 0.5122% | 0.4446% | 0.5172% | 0.6339% |
| 2 | 0.3941% | 1.0508% | 0.7772% | 1.0227% | 1.1888% |
| 3 | 0.4107% | 1.4004% | 1.0030% | 1.3766% | 1.5256% |
| 4 | 0.4275% | 1.6173% | 1.1720% | 1.6321% | 1.7454% |
| 5 | 0.4382% | 1.8171% | 1.3375% | 1.7532% | 1.9559% |
| 6 | 0.4536% | 1.9578% | 1.5143% | 1.9216% | 2.1219% |
| 7 | 0.4656% | 2.0869% | 1.6329% | 2.0530% | 2.3566% |
| 8 | 0.4700% | 2.2147% | 1.7518% | 2.3140% | 2.3260% |
| 9 | 0.4814% | 2.3307% | 1.8779% | 2.4208% | 2.3946% |
| 10 | 0.4859% | 2.4205% | 2.1141% | 2.5117% | 2.4770% |
| 11 | 0.4934% | 2.5085% | 2.2434% | 2.5222% | 2.5982% |
| 12 | 0.5025% | 2.6358% | 2.3544% | 2.5296% | 2.6936% |
| 13 | 0.4965% | 2.6834% | 2.5284% | 2.5505% | 2.8381% |
| 14 | 0.5026% | 2.8844% | 2.6226% | 2.6936% | 2.9277% |
| 15 | 0.5163% | 2.9492% | 2.7450% | 2.7256% | 3.0386% |
| 16 | 0.5144% | 3.0010% | 2.8394% | 2.7333% | 3.1154% |
| 17 | 0.5132% | 3.1150% | 2.9283% | 2.7811% | 3.3288% |
| 18 | 0.5163% | 3.1104% | 2.8843% | 2.8506% | 3.3299% |
| 19 | 0.5202% | 3.1732% | 2.9197% | 2.8456% | 3.3460% |
| 20 | 0.5225% | 3.1664% | 2.9832% | 2.8752% | 3.4739% |
| 21 | 0.5312% | 3.2085% | 3.0874% | 2.8769% | 3.5119% |
| 22 | 0.5185% | 3.2018% | 3.0088% | 2.8386% | 3.4865% |
| 23 | 0.5196% | 3.2039% | 3.0125% | 2.8614% | 3.4708% |
| 24 | 0.5229% | 3.2503% | 2.9469% | 2.9142% | 3.4229% |
| 25 | 0.5243% | 3.2280% | 2.9580% | 2.9476% | 3.3584% |
| 26 | 0.5312% | 3.2534% | 2.8180% | 2.8599% | 3.3339% |
| 27 | 0.5300% | 3.3521% | 2.8060% | 2.8813% | 3.3269% |
| 28 | 0.5266% | 3.3503% | 2.7924% | 2.8237% | 3.2029% |
| 29 | 0.5308% | 3.2479% | 2.7041% | 2.7992% | 3.1025% |
| 30 | 0.5324% | 3.2854% | 2.8646% | 2.7911% | 3.2368% |
| 31 | 0.5314% | 3.3352% | 2.8725% | 2.8038% | 3.3111% |
| 32 | 0.5344% | 3.3065% | 2.8172% | 2.7905% | 3.1914% |
| 33 | 0.5218% | 3.3277% | 2.8266% | 2.8606% | 3.1448% |
| 34 | 0.5185% | 3.3985% | 2.8025% | 2.8653% | 3.1572% |
| 35 | 0.5227% | 3.4054% | 2.7640% | 2.8847% | 3.1892% |
| 36 | 0.5183% | 3.4062% | 2.7213% | 2.9238% | 3.0652% |
| 37 | 0.5181% | 3.4343% | 2.9496% | 2.9586% | 3.2323% |
| 38 | 0.5225% | 3.5039% | 2.8014% | 3.0243% | 3.2019% |
| 39 | 0.5284% | 3.4948% | 2.8166% | 3.0330% | 3.2464% |
| 40 | 0.5267% | 3.4981% | 2.7906% | 3.0686% | 3.1959% |
| 41 | 0.5171% | 3.5703% | 2.8402% | 3.1546% | 3.2724% |
| 42 | 0.5250% | 3.5607% | 2.8864% | 3.2141% | 3.2500% |
| 43 | 0.5179% | 3.5173% | 2.9188% | 3.1705% | 3.3502% |
| 44 | 0.5114% | 3.5584% | 2.9520% | 3.1693% | 3.3623% |
| 45 | 0.5106% | 3.6376% | 2.8216% | 3.1867% | 3.4089% |
| 46 | 0.5023% | 3.6451% | 2.9459% | 3.2121% | 3.4624% |
| 47 | 0.5060% | 3.6554% | 3.3685% | 3.2900% | 3.5069% |
| 48 | 0.5071% | 3.6363% | 3.3702% | 3.2996% | 3.4843% |
| 49 | 0.5052% | 3.5704% | 2.7465% | 3.3662% | 3.3796% |
| 50 | 0.5134% | 3.6710% | 3.0324% | 3.3316% | 3.5541% |
| 51 | 0.5191% | 3.6492% | 3.0959% | 3.2446% | 3.5551% |
| 52 | 0.5133% | 3.5479% | 3.0844% | 3.2124% | 3.5615% |

Table 8: Results for the separate forecasts for the random forest model after feature selection and hyperparameter tuning with MAE as error measure

Again, a Wilcoxon test has been performed to check whether there is a significant difference between the results of predicting the total upsell KPI and predicting the three upsell KPIs A, B and C separately and then adding these predictions. It has been found that predicting the total upsell KPI is significantly better than adding up the separate predictions.

We will show four feature importance graphs for the random forest model, one for each upsell category for 20 weeks in advance. These graphs can be found in Figures 11, 12, 13 and 14.

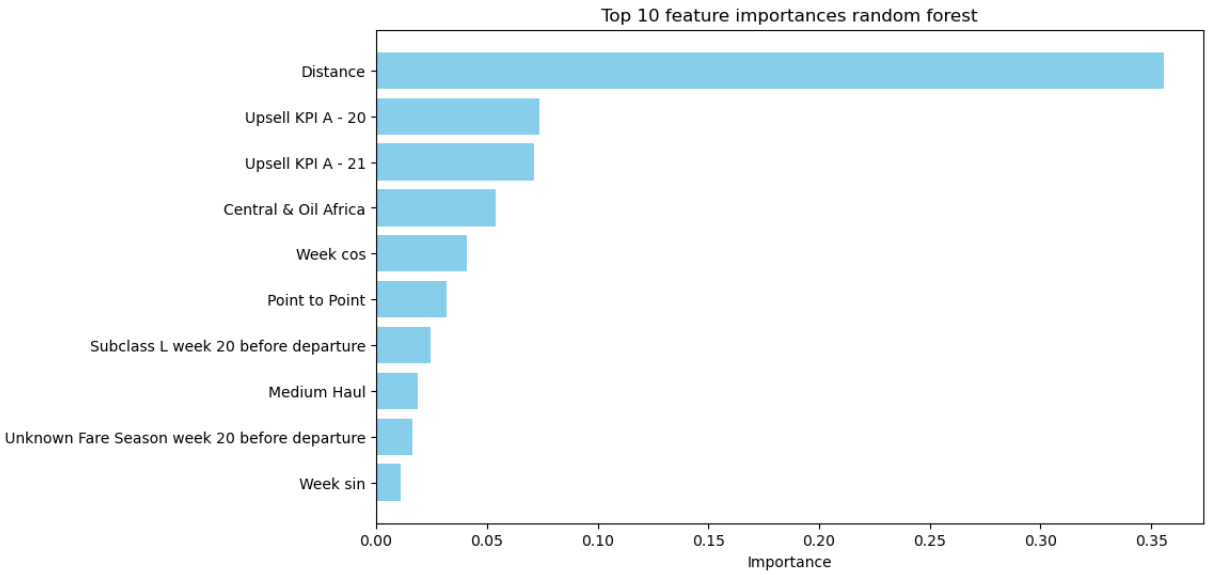


Figure 11: Feature importance for the random forest model for forecasting upsell KPI A

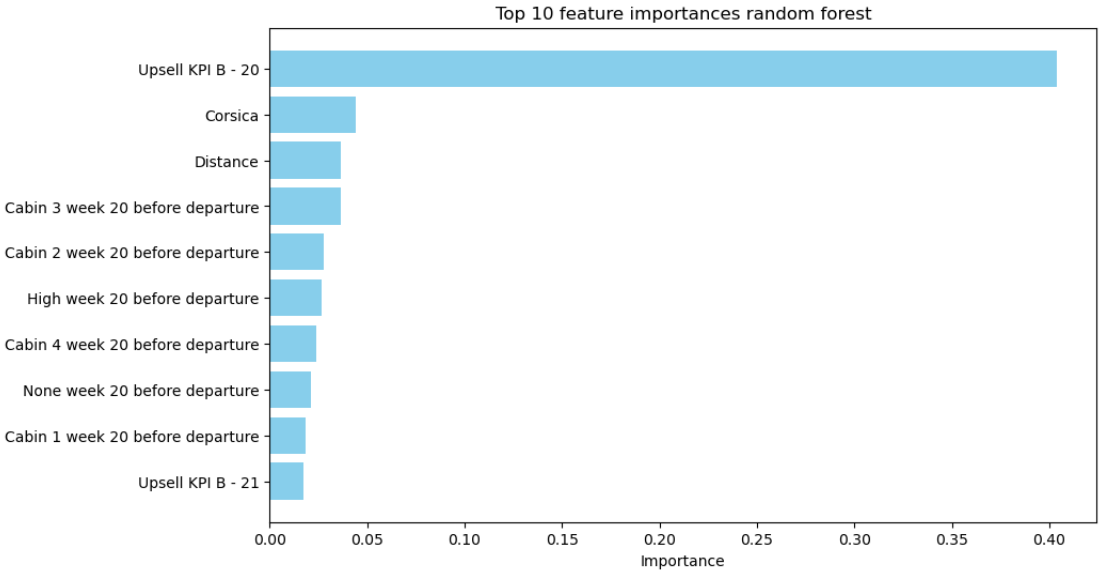


Figure 12: Feature importance for the random forest model for forecasting upsell KPI B

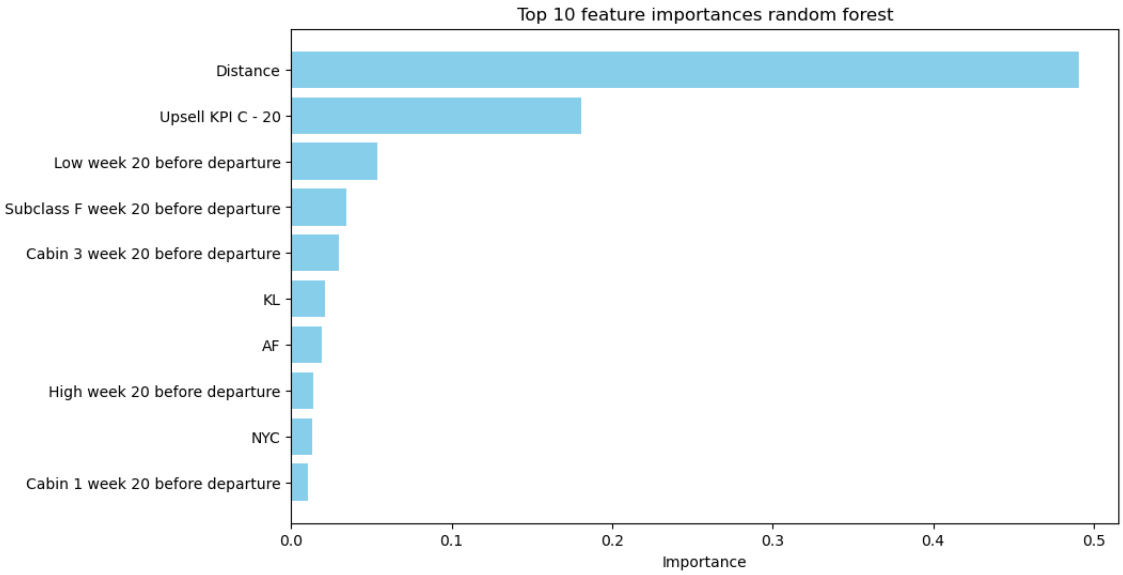


Figure 13: Feature importance for the random forest model for forecasting upsell KPI C

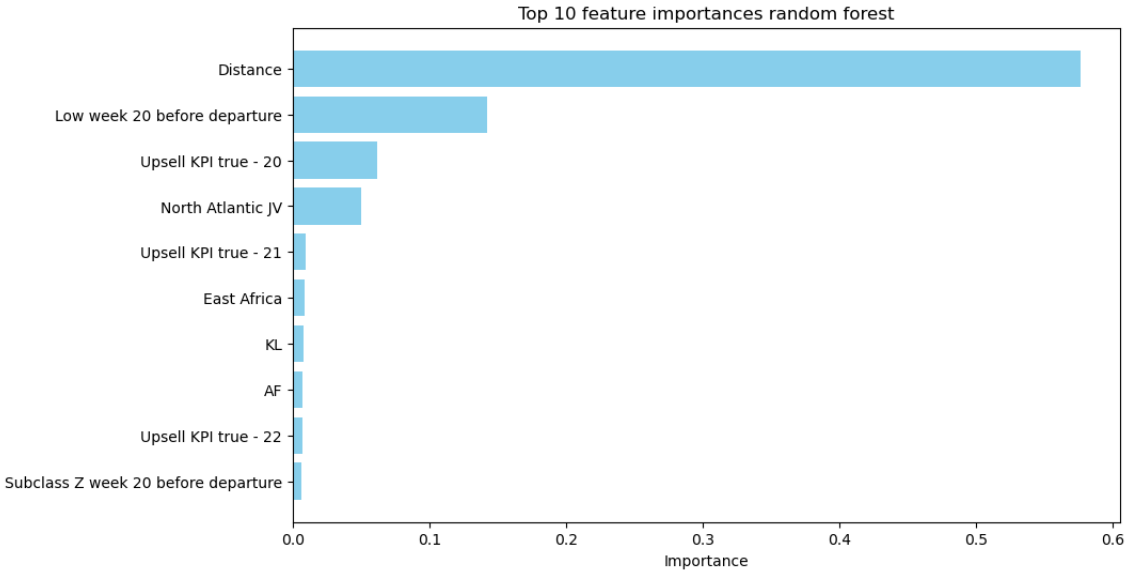


Figure 14: Feature importance for the random forest model for forecasting the total upsell KPI

We can see that distance is the most important feature for three of the models, but it is much less important to predict upsell KPI B. This could be because choosing a branded fare, for which luggage is the most important reason, is not dependent on how long your flight is, but rather on how long a passenger will be staying at their destination. In addition, we see that the corresponding upsell KPI of the previous week is in the top three features for each model.

# 6 Conclusion and discussion

This section covers the conclusion of this study, as well as its limitations and opportunities for future research. In Section 6.1, we will restate the research questions followed by the answers. In Section 6.2, the study's limitations and suggestions for future research will be addressed.

## 6.1 Conclusion

**RQ1:** What factors are important when predicting how much upsell a customer will be doing?

To answer this research questions we have first done exploratory data analysis to find if there are significant differences between the categories in a variable with respect to the upsell KPI. We can conclude that all investigated categorical variables show significant differences, and thus they can all be seen as important factors when predicting how much upsell a customer will be doing. These variables are revenue group, day of week, travel motive, corporate, age group, length of stay, frequent flyer level, booking method, subclass, farebase season, traffic type and carrier. Using the feature importance graphs from the two forecasts, it has also been found that the upsell KPI of the previous week, the distance and the week number are important.

**RQ2:** Which models are suitable to forecast the upsell KPI?

To answer this research question, we have done forecasts for predicting with and without passenger information. We have compared these forecasts to find out when the forecast with passenger information is better than the forecast without passenger information. Moreover, we have predicted the upsell KPI A, B, and C separately, as well as the total upsell KPI, and have compared predicting the upsell KPI at once with adding up the predictions for the separate upsell KPIs. We have found that, for predicting without passenger information, the XGBoost model and the random forest model can most accurately predict the upsell KPI. There is no significant difference between the performance of these models. For predicting with passenger information, the random forest model

performs significantly better than all other models. For one week until departure up to and including week 41 until departure, the model including passenger information performs better than the model without passenger information. From week 42 before departure on, the model without passenger information is more accurate and should be used. Moreover, we found that, for the forecast without passenger information, there is no significant difference between predicting the upsell KPI at once versus adding up the separate predictions for the upsell KPIs A, B and C. For the forecast with passenger information, a significant difference between these predictions has been found.

**RQ3:** Can the driving factors and forecast be used and interpreted to increase the upsell revenue?

The driving factors can be used to increase the upsell revenue by looking which variables in a group cause the highest upsell. For example, we see that the upsell KPI for the booking method indirect online is significantly lower than the upsell KPI of the other methods. This could be considered and used to think of a way to increase the upsell revenue for the indirect online booking method. However, not all variables are controllable; we cannot always change something about a variable. For example, we see that for the revenue group variable, the revenue groups intercontinental and North Atlantic JV have a higher upsell KPI. However, we cannot say that KLM and Air France should simply fly more to these destinations and should remove other destinations from their roster. This has to do with a lot of different factors, namely the slot allocations at airports, airplanes that are specifically designed for shorter routes and cannot fly longer routes, pilots that have been trained for these smaller airplanes and cabin crew that can only work on shorter routes. Furthermore, we can argue whether the link between the revenue groups and a higher upsell KPI is indeed causal. Although the revenue groups intercontinental and North Atlantic JV show a higher upsell KPI, it does not necessarily imply that increasing flights to these revenue groups would enhance the overall upsell KPI.

The forecast can be used to see what the forecasted upsell KPI is, so that Air France and KLM know what approximately their upsell KPI will be. Furthermore, the forecast can be used to increase the

upsell revenue by steering. If we see that the forecast is disappointing for a certain complex group in a certain week, we can investigate why that is the case. If the forecast is higher than expected, we should of course also look at why this is the case and if we can repeat this more often. Air France-KLM has a lot of tools in which they can visualize the data, so they can look at some of the variables that the model considers and take actions upon that.

## 6.2 Discussion

This research has focused mostly on predicting the upsell KPI on a weekly complex group level. This could be expanded in a future study by also forecasting on other levels, for example by zooming in on specific sublines or even flights. This might help analysts to see which sublines or flights have lower predictions and focus on these sublines or flights.

Furthermore, the test set is now relatively small; we only take into account the first four months of 2024 as a test set. It could be better to also use an entire year as a test set; for example two years for training and one year for testing, so that we can evaluate the models on an entire year of predictions. At the time, this was not possible, since the years before 2022 were not reliable enough due to Covid and the impact this epidemic had on the airline industry.

In addition, this research takes as upsell the complete price for a ticket as upsell when buying a branded fare for example. It might be better to only take the difference between buying ticket A, a ticket without a branded fare, and ticket B, the exact same ticket, but with a branded fare. This has not been done simply because this data is not available yet.

For future research, it would be good to take into account the specific actions that analysts take in order to increase the (upsell) revenue. For example, if analysts change the prices of a certain cabin on a certain route, it would change passenger behavior and, hopefully, increase the upsell KPI. These actions are now not taken into account directly. Thus, when a forecast is relatively low

as to what the analysts would like to see and they perform an action based on that, the forecast does not change directly based on this action. Therefore, if this could be taken into account, the forecast can predict how it would change according to an action, and the analysts can then use this information to choose the best possible action.

# References


[1] URL: https://www.klm.nl/information/corporate.

[2] URL: https://www.airfranceklm.com/en/group/purpose.

[3] David Warnock-Smith, John F. O'Connell, and Mahnaz Maleki. "An analysis of ongoing trends in airline ancillary revenues". In: *Journal of Air Transport Management* 64 (Sept. 2017), pp. 42–54. ISSN: 0969-6997. DOI: 10.1016/j.jairtraman.2017.06.023. URL: https://dx.doi.org/10.1016/j.jairtraman.2017.06.023.

[4] Brian Pearce. *The Outlook for Commercial Air Transport*. IATA World Passenger Symposium. 2014. URL: www.iata.org/economics.

[5] John F. O'Connell and David Warnock-Smith. "An investigation into traveler preferences and acceptance levels of airline ancillary revenues". In: *Journal of Air Transport Management* 33 (Oct. 2013), pp. 12–21. ISSN: 0969-6997. DOI: 10.1016/j.jairtraman.2013.06.006. URL: https://dx.doi.org/10.1016/j.jairtraman.2013.06.006.

[6] B. Wislon. *Airlines Find the Money with Ancillary Revenues*. 2014. URL: https://businessjournalism.org/2014/12/airlines-find-the-money-with-ancillary-revenues/.

[7] Mark Shaw et al. "Third party ancillary revenues in the airline sector: An exploratory study". In: *Journal of Air Transport Management* 90 (Jan. 2021), p. 101936. ISSN: 0969-6997. DOI: 10.1016/j.jairtraman.2020.101936. URL: https://dx.doi.org/10.1016/j.jairtraman.2020.101936.

[8] Woon-Kyung Song and Hyun Cheol Lee. "An analysis of traveler need for and willingness to purchase airline dynamic packaging: A Korean case study". In: *Journal of Air Transport Management* 82 (Jan. 2020), p. 101735. ISSN: 0969-6997. DOI: 10.1016/j.jairtraman.2019.101735. URL: https://dx.doi.org/10.1016/j.jairtraman.2019.101735.

[9] Richard A Davis Peter J Brockwell. *Introduction to Time Series and Forecasting*. 3rd ed. Springer, 2016.

[10] Bohdan M. Pavlyshenko. "Linear, machine learning and probabilistic approaches for time series analysis". In: *2016 IEEE First International Conference on Data Stream Mining; Processing (DSMP)* (Aug. 2016). DOI: 10.1109/dsmp.2016.7583582. URL: https://dx.doi.org/10.1109/dsmp.2016.7583582.

[11] David J. Olive. *Linear Regression*. Springer International Publishing, 2017. DOI: 10.1007/978-3-319-55252-1.

[12] Yan-yan Song and Ying Lu. "Decision tree methods: applications for classification and prediction". In: *Shanghai archives of psychiatry* 27.2 (2015), pp. 130–135. DOI: https://doi.org/10.11919/j.issn.1002-0829.215044.

[13] Leo Breiman. "Random Forests". In: *Machine Learning* 45.1 (2001), pp. 5–32.

[14] Tianqi Chen and Carlos Guestrin. "XGBoost". In: *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (2016). DOI: 10.1145/2939672.2939785.

[15] William Remus and Marcus O'Connor. "Neural Networks for Time-series forecasting". In: *International Series in Operations Research; Management Science* (2001), pp. 245–256. DOI: 10.1007/978-0-306-47630-3_12.

[16] Sunder Kekre, Thomas E. Morton, and Timothy L. Smunt. "Forecasting using partially known demands". In: *International Journal of Forecasting* 6 (Jan. 1990), pp. 115–125. ISSN: 0169-2070. DOI: 10.1016/0169-2070(90)90102-h. URL: https://dx.doi.org/10.1016/0169-2070(90)90102-h.

[17] Vladimir Kurbalija, Mirjana Ivanović, and Zoran Budimac. "Case-based curve behaviour prediction". In: *Software: Practice and Experience* 39 (July 2008), pp. 81–103. ISSN: 0038-0644, 1097-024X. DOI: 10.1002/spe.891. URL: https://dx.doi.org/10.1002/spe.891.

[18] Yanfeng Zhou et al. "Willingness to pay for economy class seat selection: From a Chinese air consumer perspective". In: *Research in Transportation Business; Management* 37 (Dec. 2020), p. 100486. ISSN: 2210-5395. DOI: 10.1016/j.rtbm.2020.100486. URL: https://dx.doi.org/10.1016/j.rtbm.2020.100486.

[19] Paul Chiambaretto. "Air passengers' willingness to pay for ancillary services on long-haul flights". In: *Transportation Research Part E: Logistics and Transportation Review* 147 (Mar. 2021), p. 102234. ISSN: 1366-5545. DOI: 10.1016/j.tre.2021.102234. URL: https://dx.doi.org/10.1016/j.tre.2021.102234.

[20] Saravanan Thirumuruganathan et al. "Will they take this offer? A machine learning price elasticity model for predicting upselling acceptance of premium airline seating". In: *Information; Management* 60 (Apr. 2023), p. 103759. ISSN: 0378-7206. DOI: 10.1016/j.im.2023.103759. URL: https://dx.doi.org/10.1016/j.im.2023.103759.

[21] Noora Al Emadi et al. "Will You Buy It Now?: Predicting Passengers that Purchase Premium Promotions Using the PAX Model". In: *Journal of Smart Tourism* 1 (Mar. 2021), pp. 53–64. ISSN: 2765-2157. DOI: 10.52255/smarttourism.2021.1.1.7. URL: https://dx.doi.org/10.52255/smarttourism.2021.1.1.7.

[22] Thanavathi C. *Advanced Educational Research and Statistics*. Dec. 2017.

[23] Eva Ostertagova, Oskar Ostertag, and Jozef Kováč. "Methodology and Application of the Kruskal-Wallis Test". In: *Applied Mechanics and Materials* 611 (Aug. 2014), pp. 115–120.

[24] Dario Radečić. *How to handle cyclical data in machine learning*. URL: https://betterdatascience.com/cyclical-data-machine-learning/.

[25] Rezhna H. Faraj Peshawa J. Muhammad Ali. "Data Normalization and Standardization: A Technical Report". In: *Machine Learning Technical Reports* 1.1 (2014), pp. 1–6.

[26] Jorge R. Vergara and Pablo A. Estévez. "A review of feature selection methods based on mutual information". In: *Neural Computing and Applications* 24 (Mar. 2013), pp. 175–186. ISSN: 0941-0643, 1433-3058. DOI: 10.1007/s00521-013-1368-0. URL: https://dx.doi.org/10.1007/s00521-013-1368-0.

[27] Daniel Berrar. "Cross-Validation". In: *Encyclopedia of Bioinformatics and Computational Biology* (2019), pp. 542–545. DOI: 10.1016/b978-0-12-809633-8.20349-x. URL: https://dx.doi.org/10.1016/b978-0-12-809633-8.20349-x.

[28] Rith Pansanga. "Optimizing XGBoost: A Guide to Hyperparameter Tuning". In: *Medium* (2023). URL: https://medium.com/@rithpansanga/optimizing-xgboost-a-guide-to-hyperparameter-tuning-77b6e48e289d.

[29] Frank Wilcoxon. "Individual Comparisons by Ranking Methods". In: *Biometrics Bulletin* 1 (Dec. 1945), p. 80. ISSN: 0099-4987. DOI: 10.2307/3001968. URL: https://dx.doi.org/10.2307/3001968.

# A Appendix

## A.1 Data Analysis

### A.1.1 Revenue group

A box plot of the different revenue groups versus the upsell KPI can be seen in figure 15. We see that there seems to be quite a difference between the revenue groups for the upsell KPI. To verify this, histograms and Q-Q plots are plotted to check for normality. These plots can be found in figures 16 and 17. The data seems to be normally distributed. The Kolmogorov-Smirnov test has also been performed to check for normality, which also shows that the data is normally distributed. Therefore, the one-way ANOVA test is used, and it is found that the differences between the revenue groups are significant.

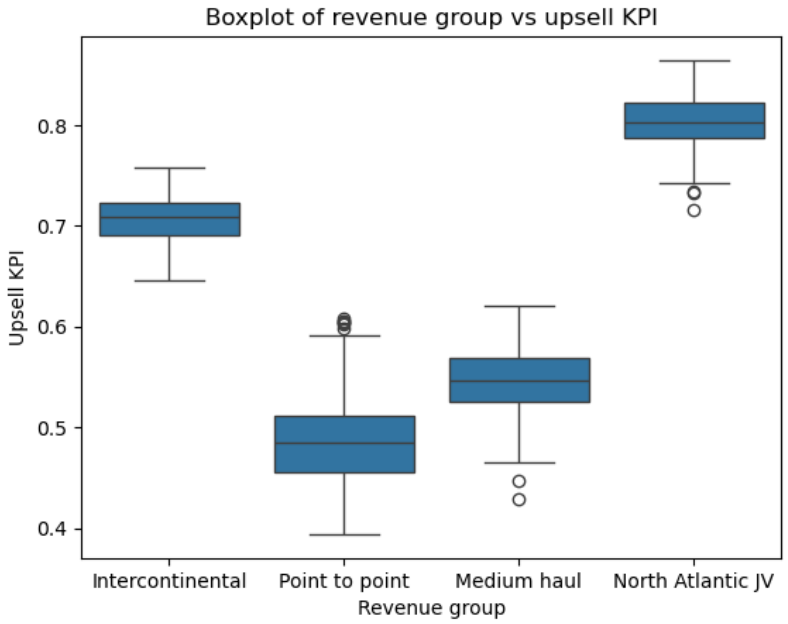


Figure 15: Box plot for the revenue groups

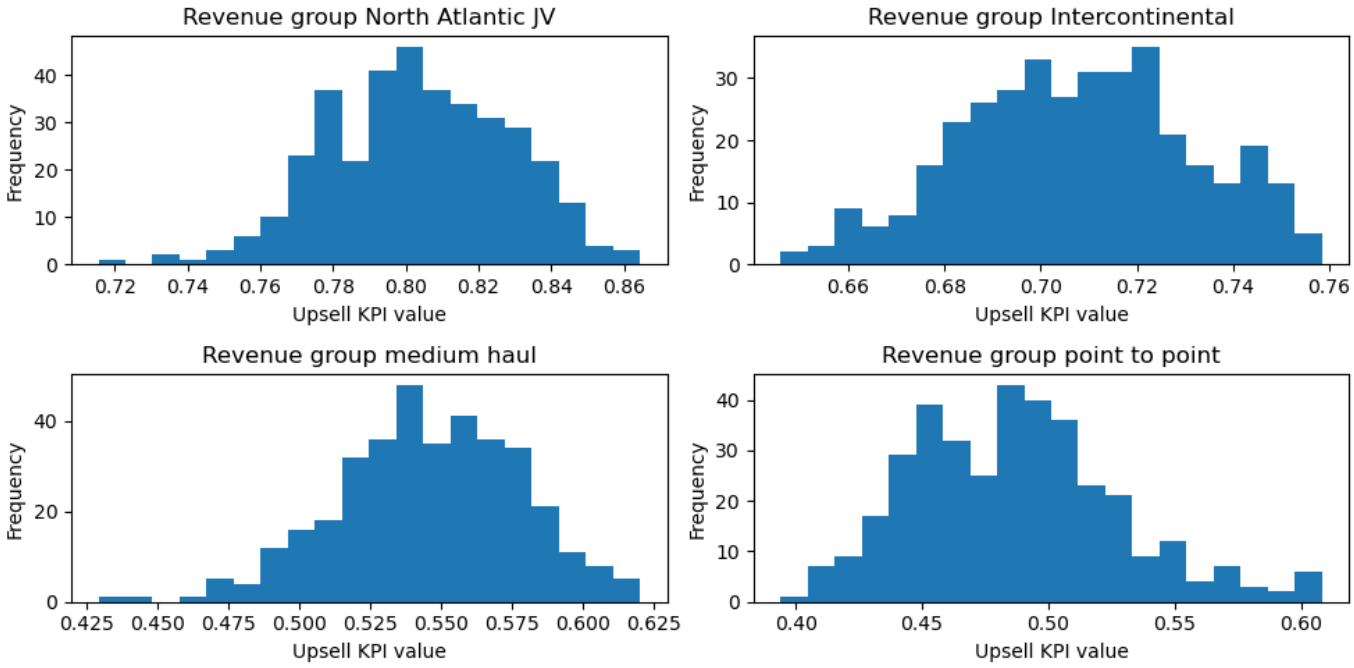


Figure 16: Histograms for the revenue groups

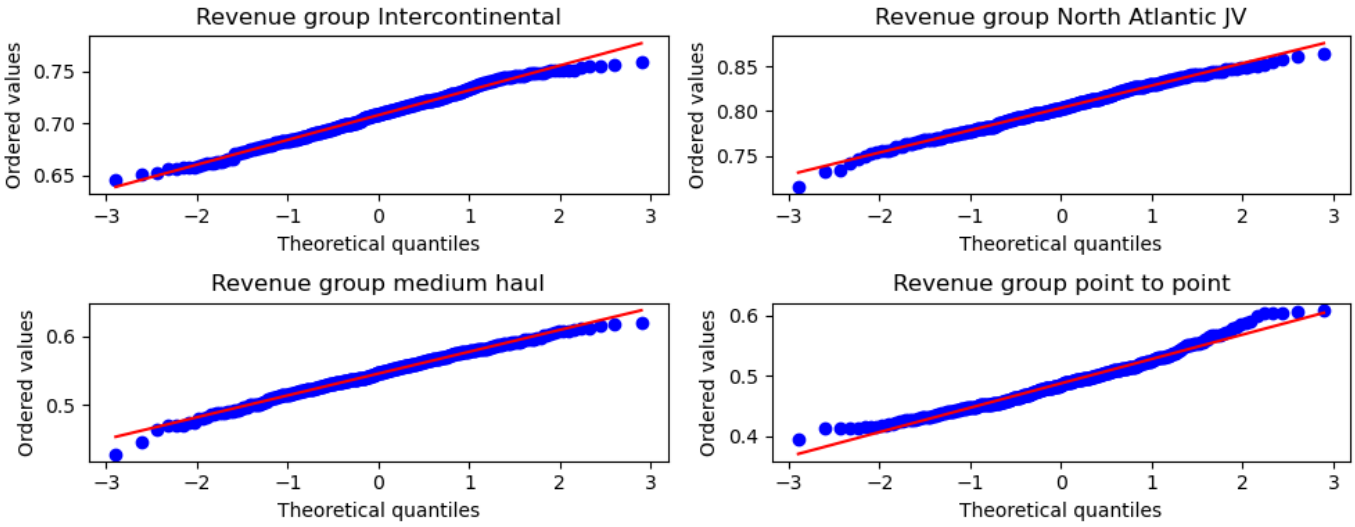


Figure 17: Q-Q plots for the revenue groups

### A.1.2 Carrier

A box plot of the different carriers and their corresponding upsell KPI can be found in Figure 18. We can see that Air France has a higher upsell KPI than KLM.

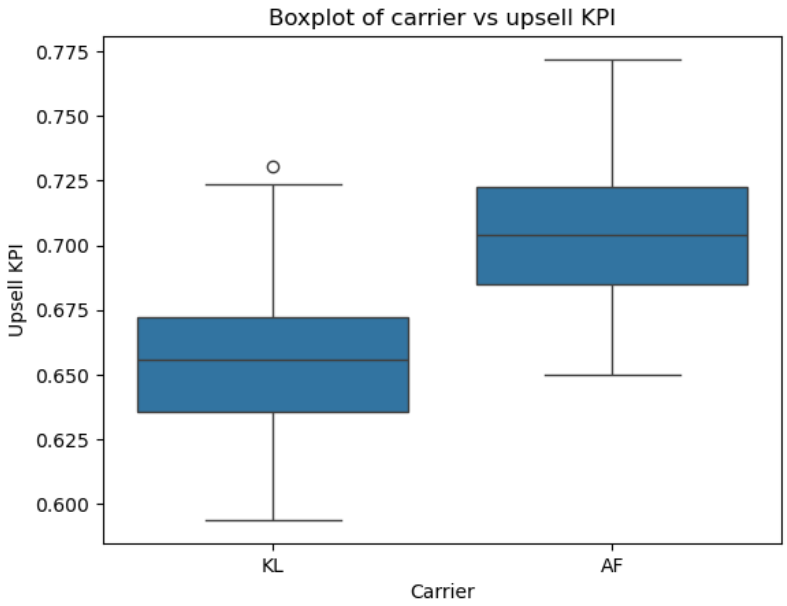


Figure 18: Boxplot for the carrier variable

We have made histograms and Q-Q plots to check for normality. These can be found in Figures 19 and 20. The data seems to have a normal distribution, although the histograms are inconclusive. Therefore, Kolmogorov-Smirnov tests have been performed to verify normality. It has been found that both the data for KLM and Air France are normally distributed. Consequently, the one-way ANOVA test has been done to find if there is a significant difference between the carriers KLM and Air France. A p-value of 6.39e-102 has been found and thus it has been concluded that the different carriers have significant differences between them.

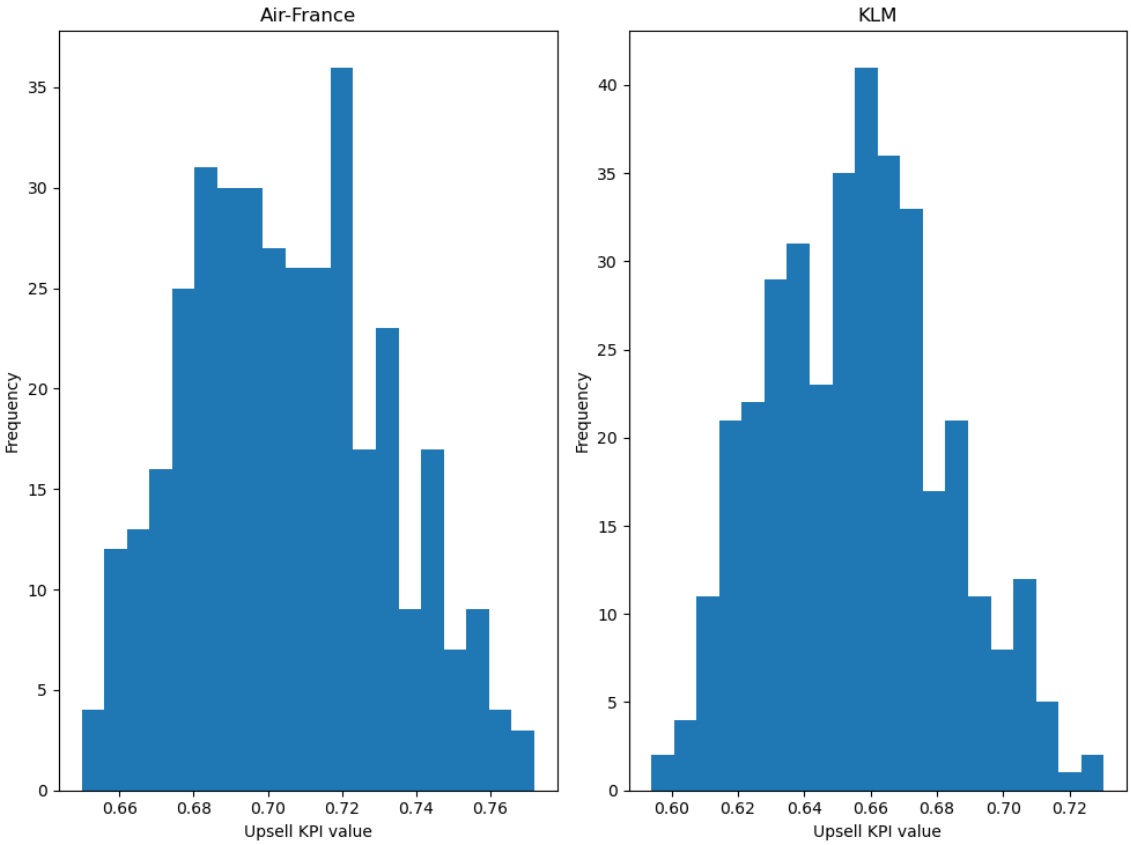


Figure 19: Histograms for the carriers

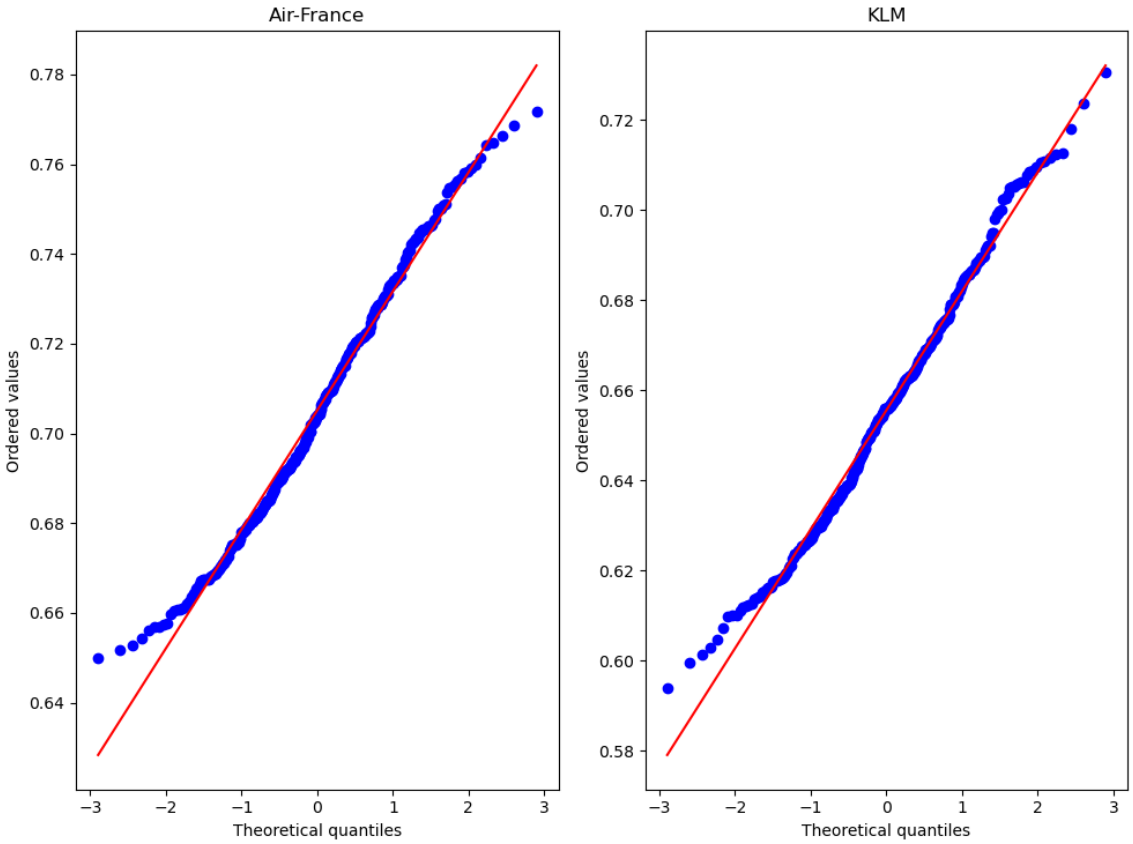


Figure 20: Q-Q plots for the carriers

### A.1.3 Day of week

A box plot of the different days of the week and their corresponding upsell KPI can be found in figure 21. It can be seen that there are quite some differences between the days of the week, especially between the weekdays (Monday, Tuesday, Wednesday, and Thursday) and the weekend days (Friday, Saturday, and Sunday).

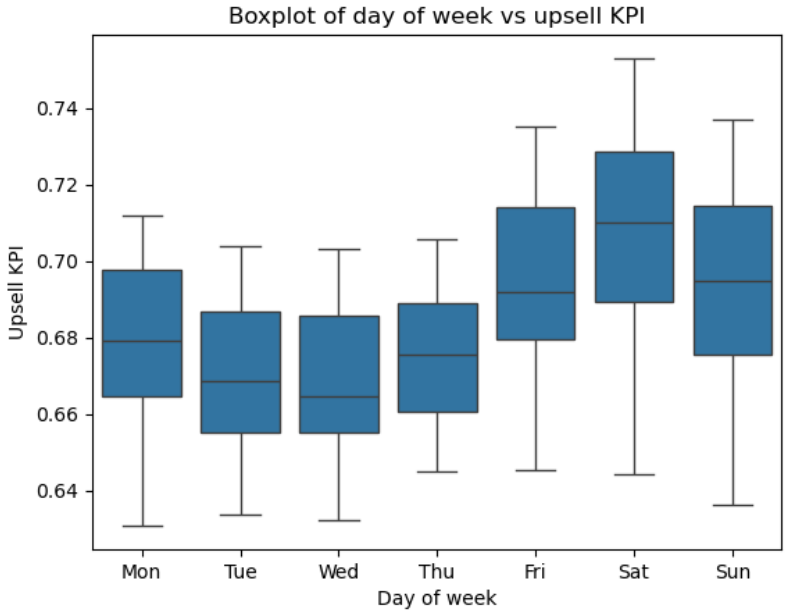


Figure 21: Box plot for the day of week variable

Histograms and Q-Q plots have been made to verify normality. These can be found in figures 22 and 23. It appears that the data follows a normal distribution, although there is some uncertainty. The Kolmogorov-Smirnov tests have been performed to check for normality for all days of the week, and all found p-values are below 0.05. Therefore, the data is normally distributed. Thus, the one-way ANOVA test has been executed to find if there is a significant difference between the variables. A p-value of 8.96e-23 has been found and thus it has been concluded that the different categories have significant differences between them.

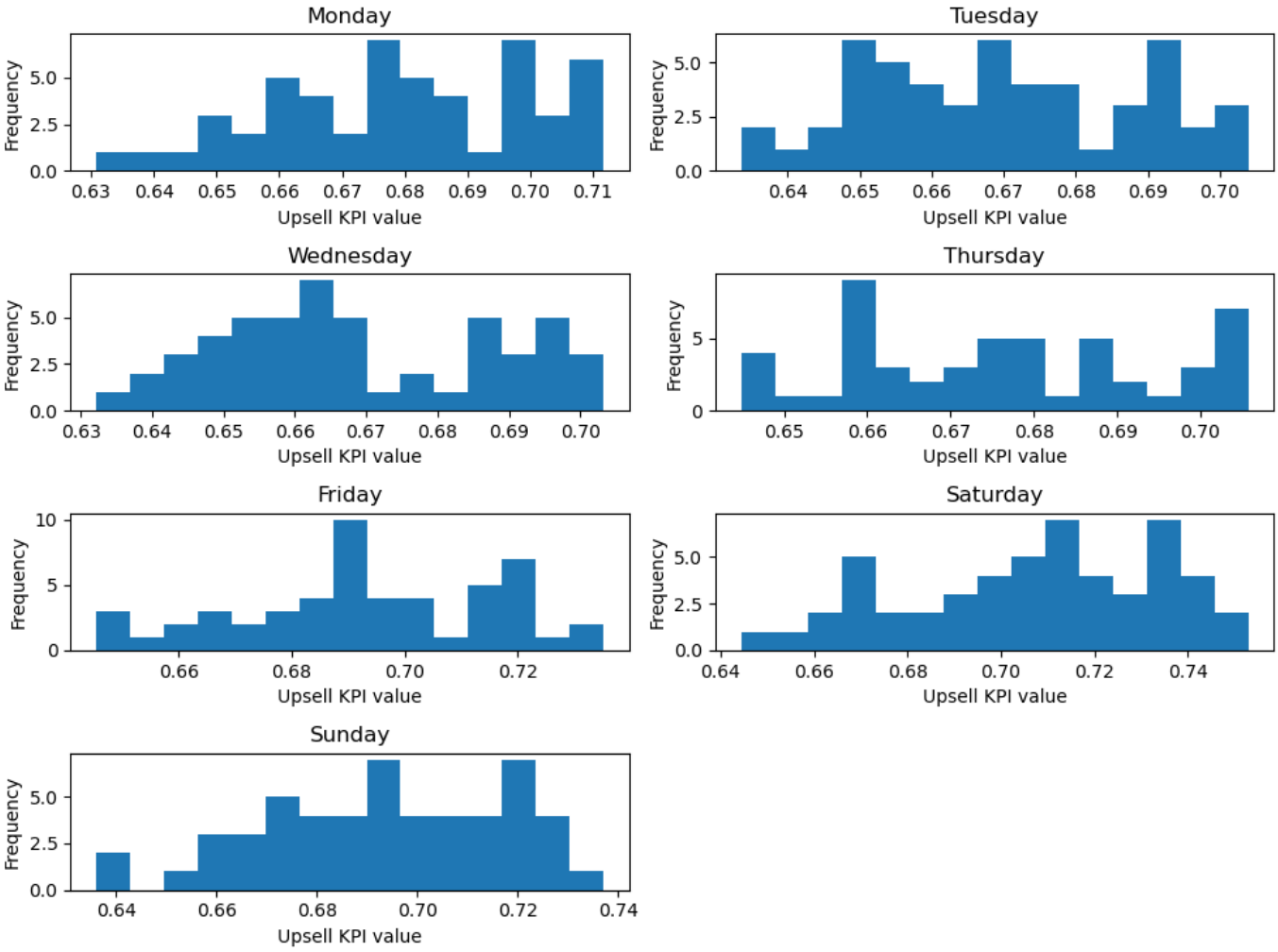


Figure 22: Histograms for the day of week variable

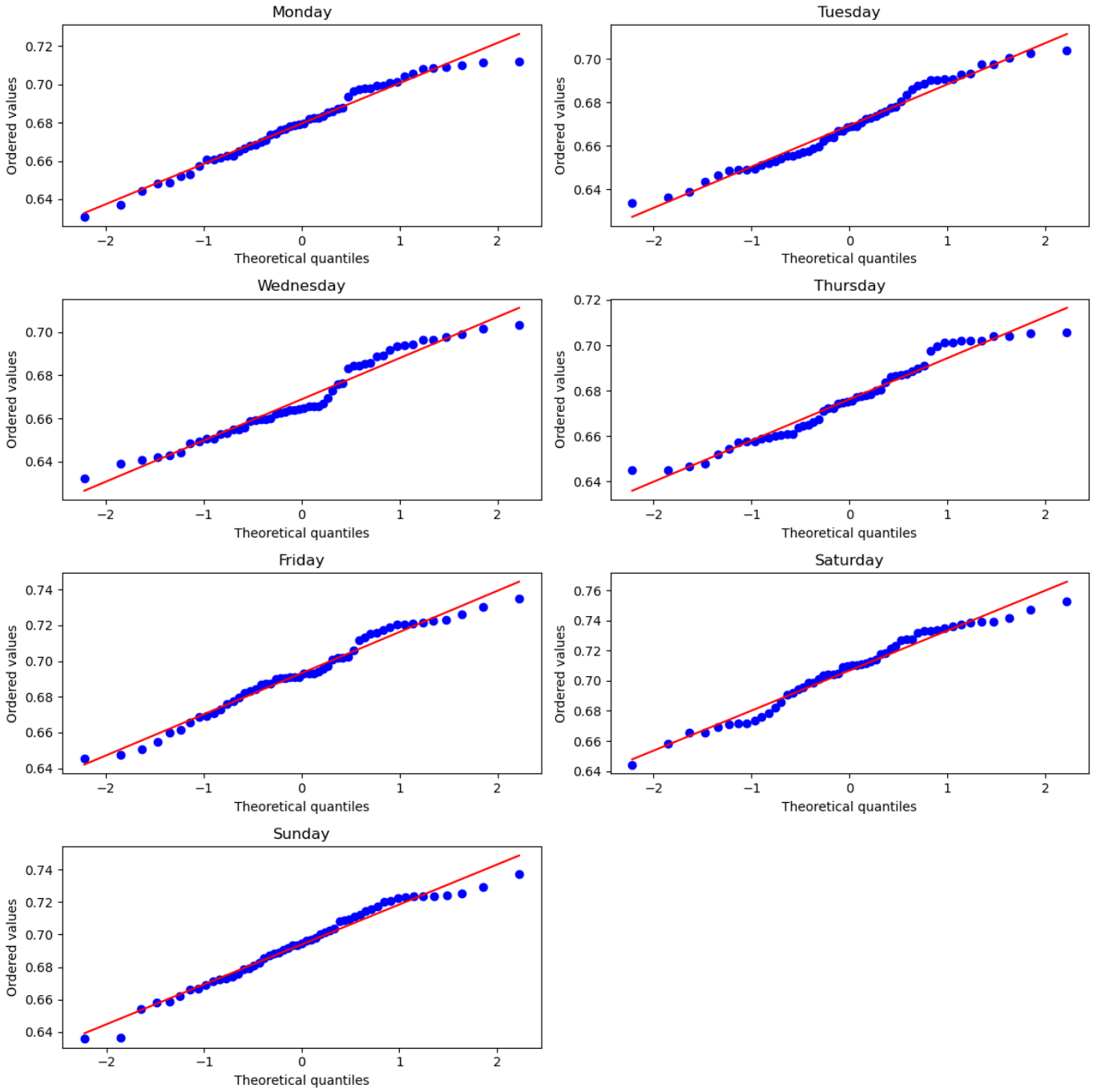


Figure 23: Q-Q plots for the day of week variable

### A.1.4 Travel motive

A box plot was created for the different travel motives. The plot, which can be found in figure 24, suggests a difference between the travel motives, but it is not apparent.

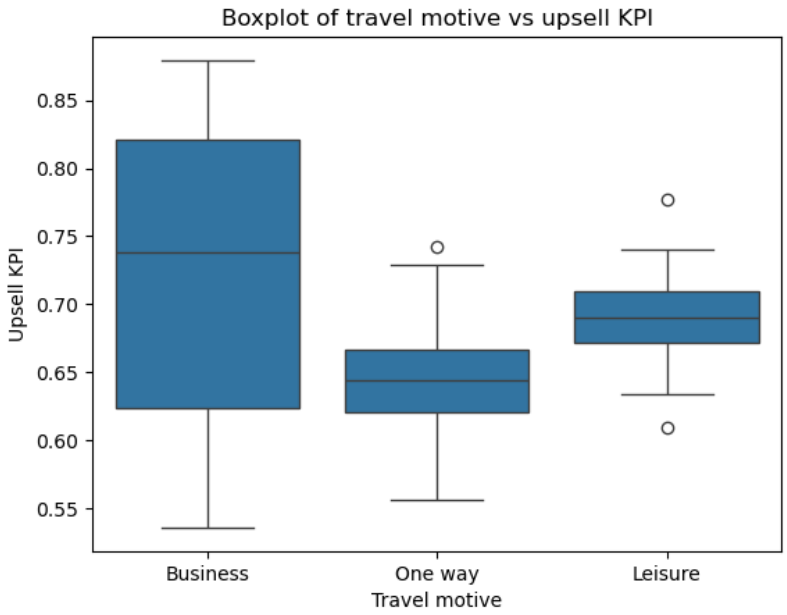


Figure 24: Box plot for travel motive

Histograms and Q-Q plots have been made to check for normality. These plots can be found in figures 25 and 26. We can see that for the travel motives one way and leisure, the data is normally distributed. However, for the travel motive business, the data is not normally distributed as we can see. Therefore, the Kruskal-Wallis test has been performed to check for significant differences. A p-value of 6.51e-58 has been found and therefore it is concluded that significant differences exist between the three travel motives.

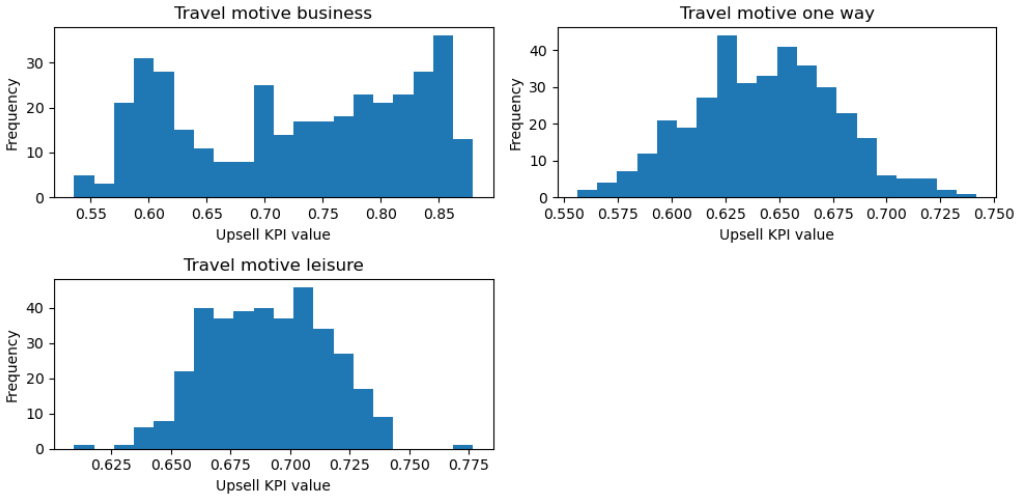


Figure 25: Histograms for the different travel motives

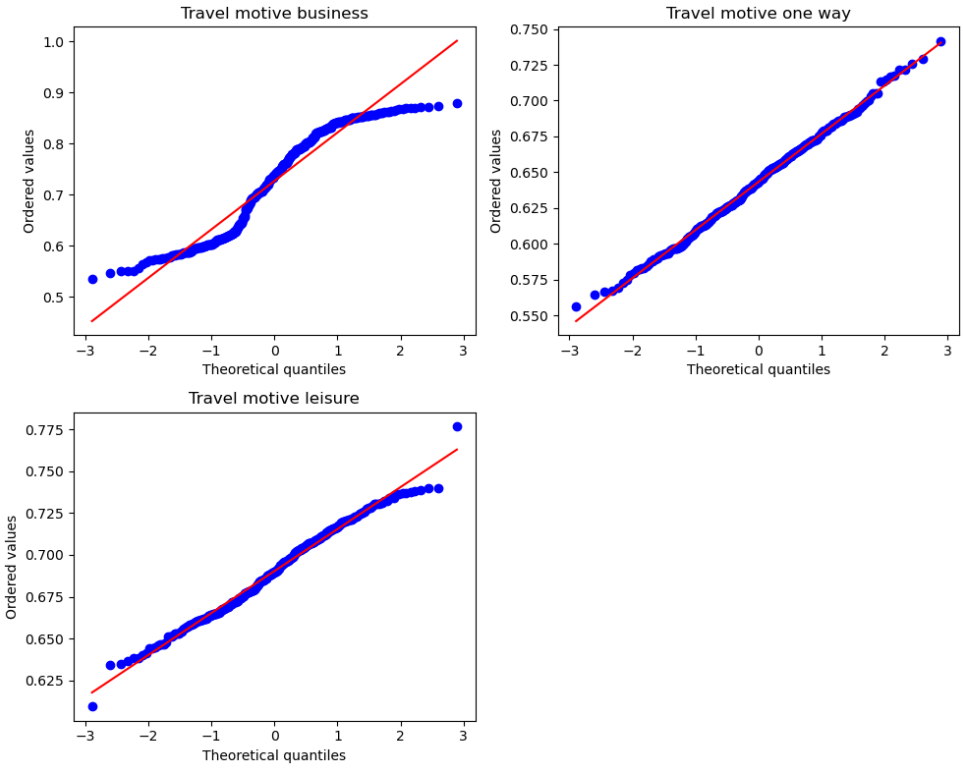


Figure 26: Q-Q plots for the different travel motives

### A.1.5 Corporate

The corporate variable is a binary variable, with corporate and non-corporate as categories. A box plot has been made and can be found in figure 27. We see that the upsell KPI for corporate passengers has a larger variability and seems to be higher on average than for non-corporate passengers.

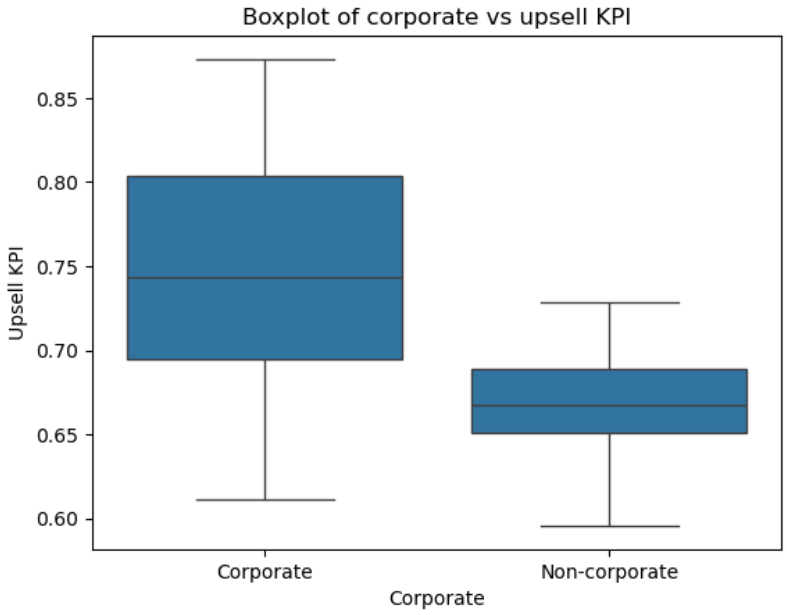


Figure 27: Box plot for corporate

Histograms and Q-Q plots have been made for corporate and non-corporate passengers. They can be found in figures 28 and 29. Both seem to be normally distributed. The Kolmogorov-Smirnov test has also been performed to check for normality and since both p-values are below 0.05, it is concluded that both variables come from a normal distribution. Therefore, a one-way ANOVA test has been conducted to check for significant differences. A p-value of 9.09e-88 has been found and thus, the differences between the two groups of passengers are significant.

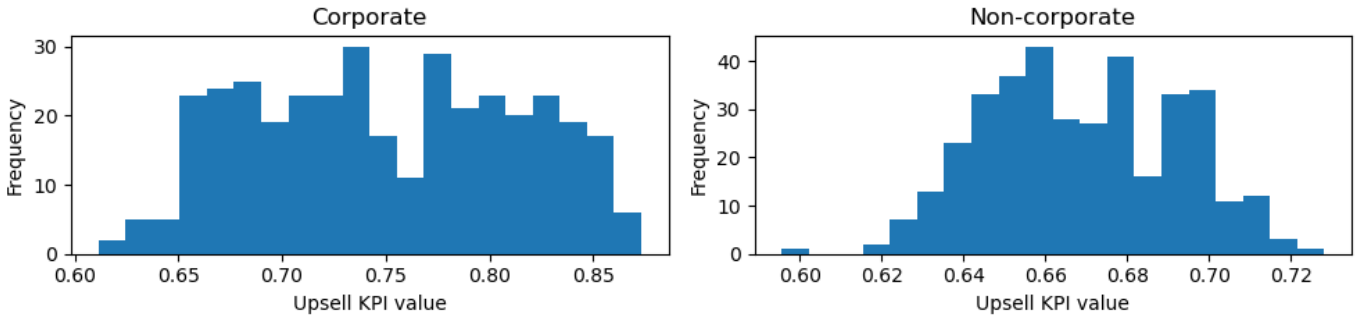


Figure 28: Histograms for corporate and non-corporate passengers

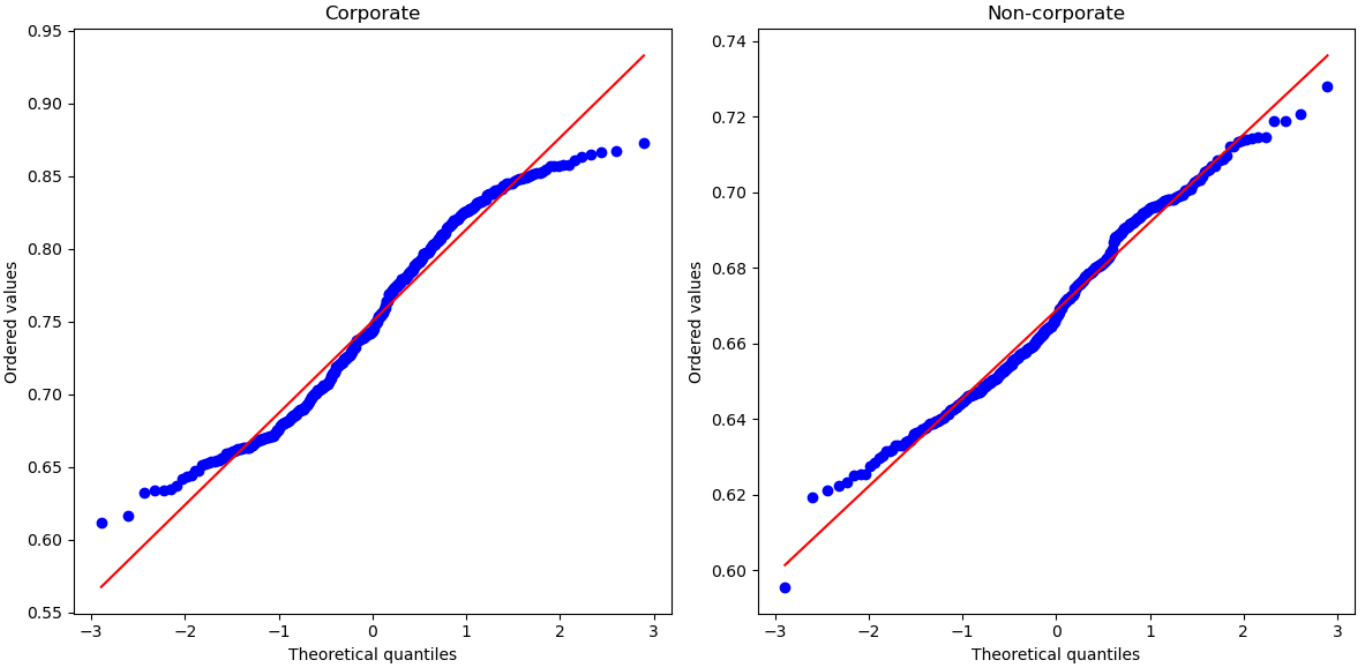


Figure 29: Q-Q plots for corporate and non-corporate passengers

### A.1.6 Age group

The variable age group consists of three categories: adult, child, and infant. A box plot of these three categories and their respective upsell KPIs can be found in figure 30. As can be seen in the plot, the age group child seems to have a substantial difference from the other two age groups.

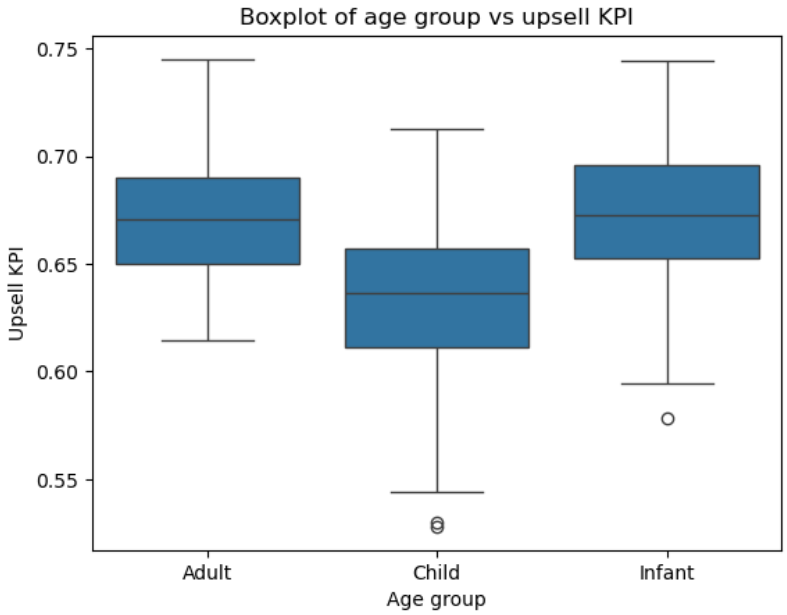


Figure 30: Box plot for age group

For all three categories, histograms and Q-Q plots have been made. As can be seen in figures 31 and 32, the data appears to be normally distributed. The Kolmogorov-Smirnov test also indicates that the data is normally distributed, since all p-values are below 0.05. Consequently, a one-way ANOVA test has been conducted to see if there are significant differences between age groups. A p-value of 1.2e-70 has been found and thus, it has been concluded that there is a significant difference between the age groups.

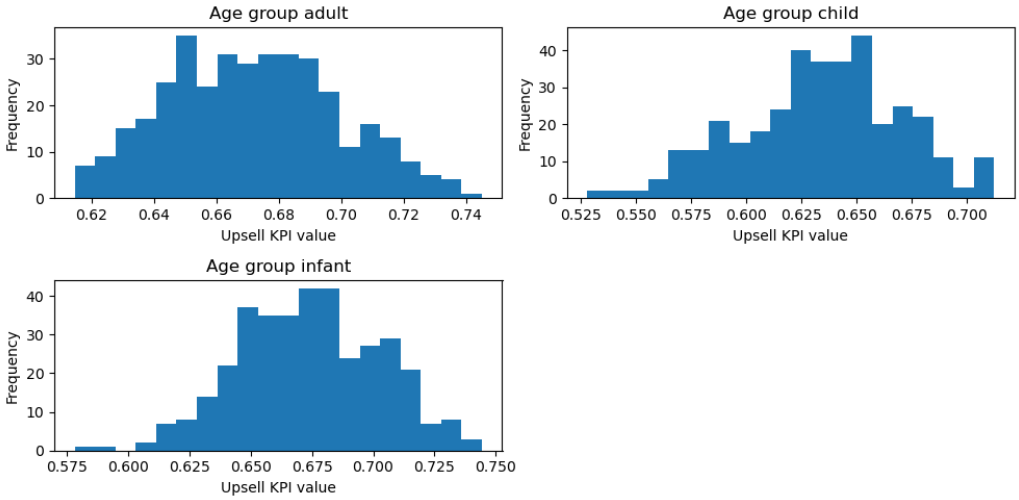


Figure 31: Histograms for age group

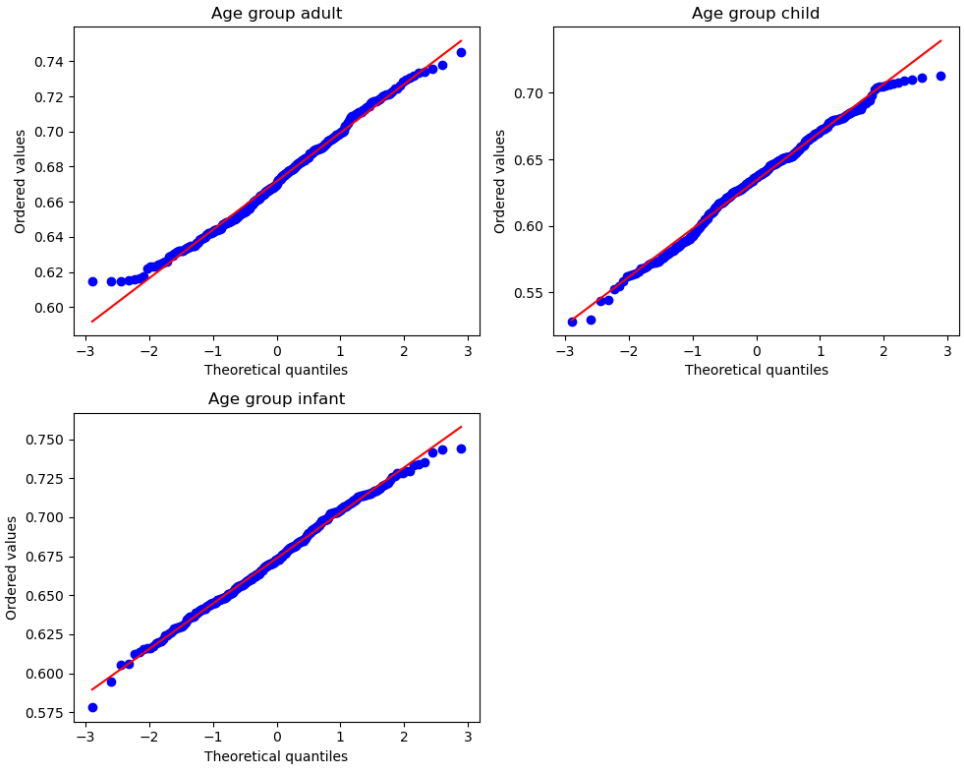


Figure 32: Q-Q plots for age group

### A.1.7 Length of stay

The length of stay variable consists of nine categories: zero days, one day, two days, three days, four days, five days, six days, seven days and more, and unknown. The unknown category consists of passengers who have booked a one-way ticket and some passengers of whom not enough data is present to know their length of stay. A box plot of this variable can be found in figure 33. We can see that there are big differences between the upsell KPI for the different length of stay categories.

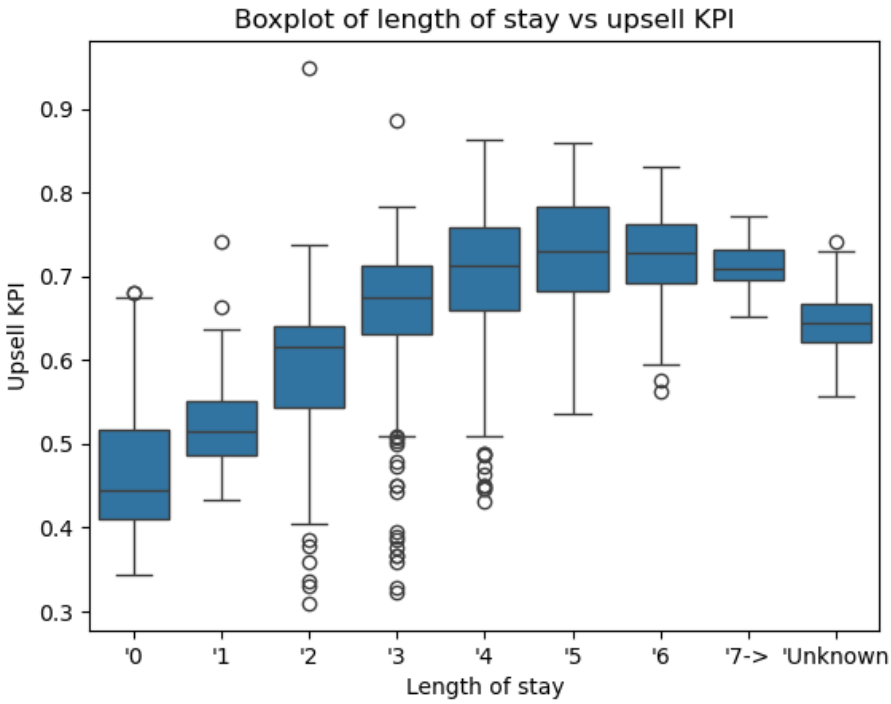


Figure 33: Box plot for length of stay

To check for normality, histograms and Q-Q plots have been made for each category of length of stay. These graphs can be found in figures 34 and 35. The Q-Q plots show some deviations, and thus the Kolmogorov-Smirnov test is performed to verify that the data is normally distributed. All p-values of this test are below 0.05, and thus it is concluded that the data for all length of stay categories is normally distributed. A one-way ANOVA test has been performed to check whether

the differences between the categories are significant. Since a p-value below 0.05 has been found, it is concluded that the differences are significant.

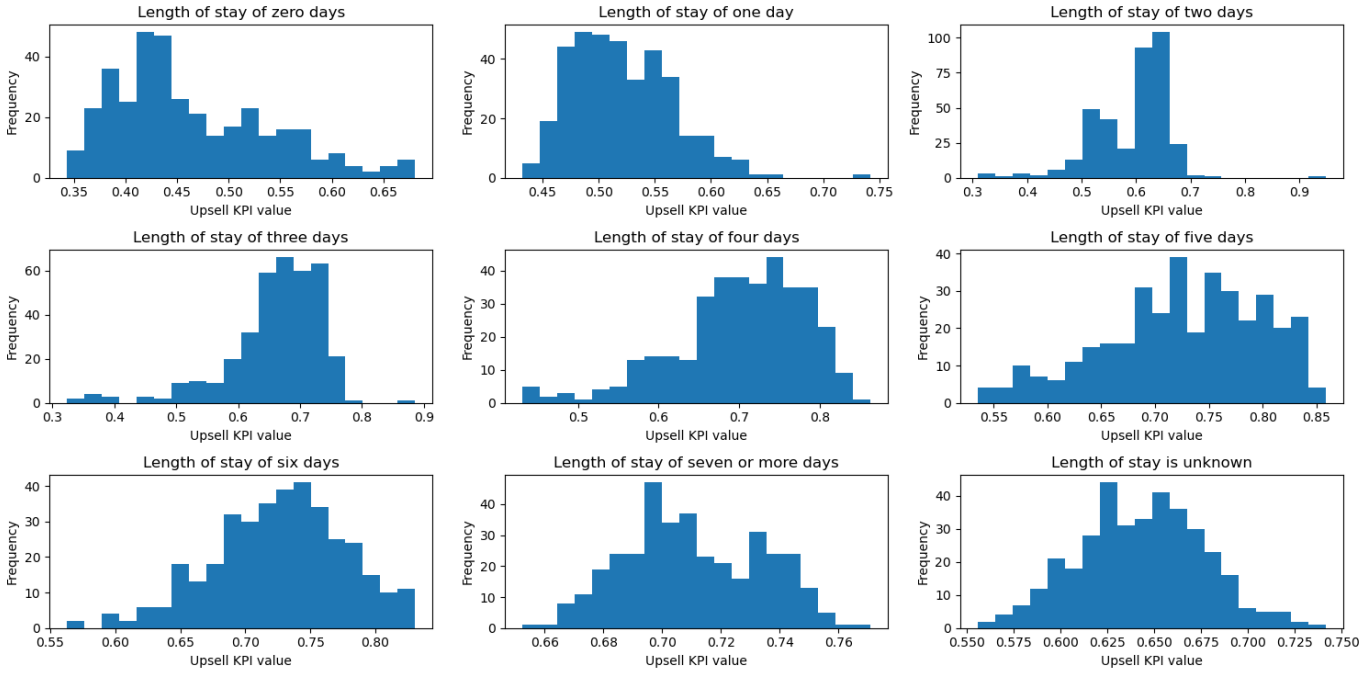


Figure 34: Histograms for length of stay

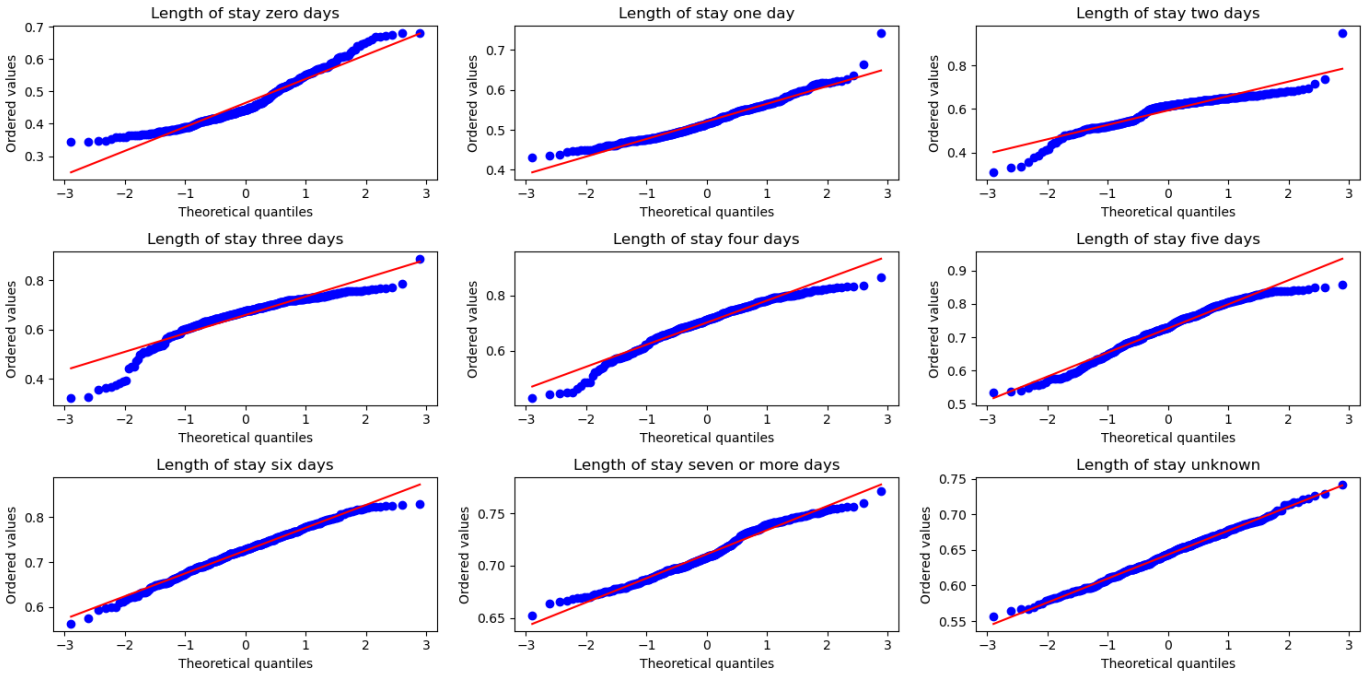


Figure 35: Q-Q plots for length of stay

### A.1.8 Frequent flyer level

The frequent flyer level variable consists of six categories: the four frequent flyer levels explorer, silver, gold, and platinum, the non-frequent flyers, and the passengers with unknown frequent flyer level. For these categories, a box plot has been made, which can be seen in figure 36. It can be seen that the upsell KPI for frequent flyer levels seems to be higher per level upward, and also than the non-frequent flyer and unknown frequent flyer level.

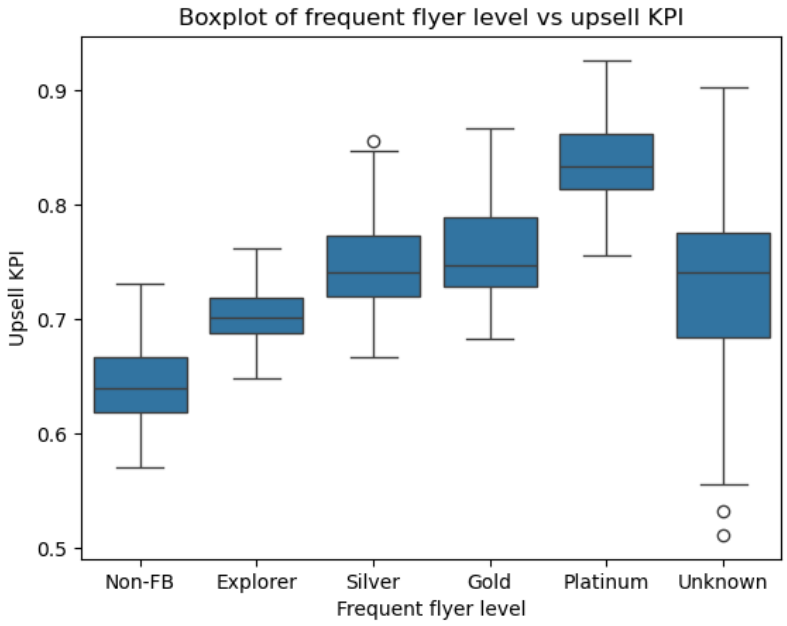


Figure 36: Box plot for frequent flyer level

To check for normality, histograms and Q-Q plots have been made for the different frequent flyer levels. They can be found in figures 37 and 38. The data seems to be normally distributed according to these graphs and the Kolmogorov-Smirnov test confirms this. Accordingly, a one-way ANOVA test has been performed and it is found that there are significant differences between the frequent flyer levels.

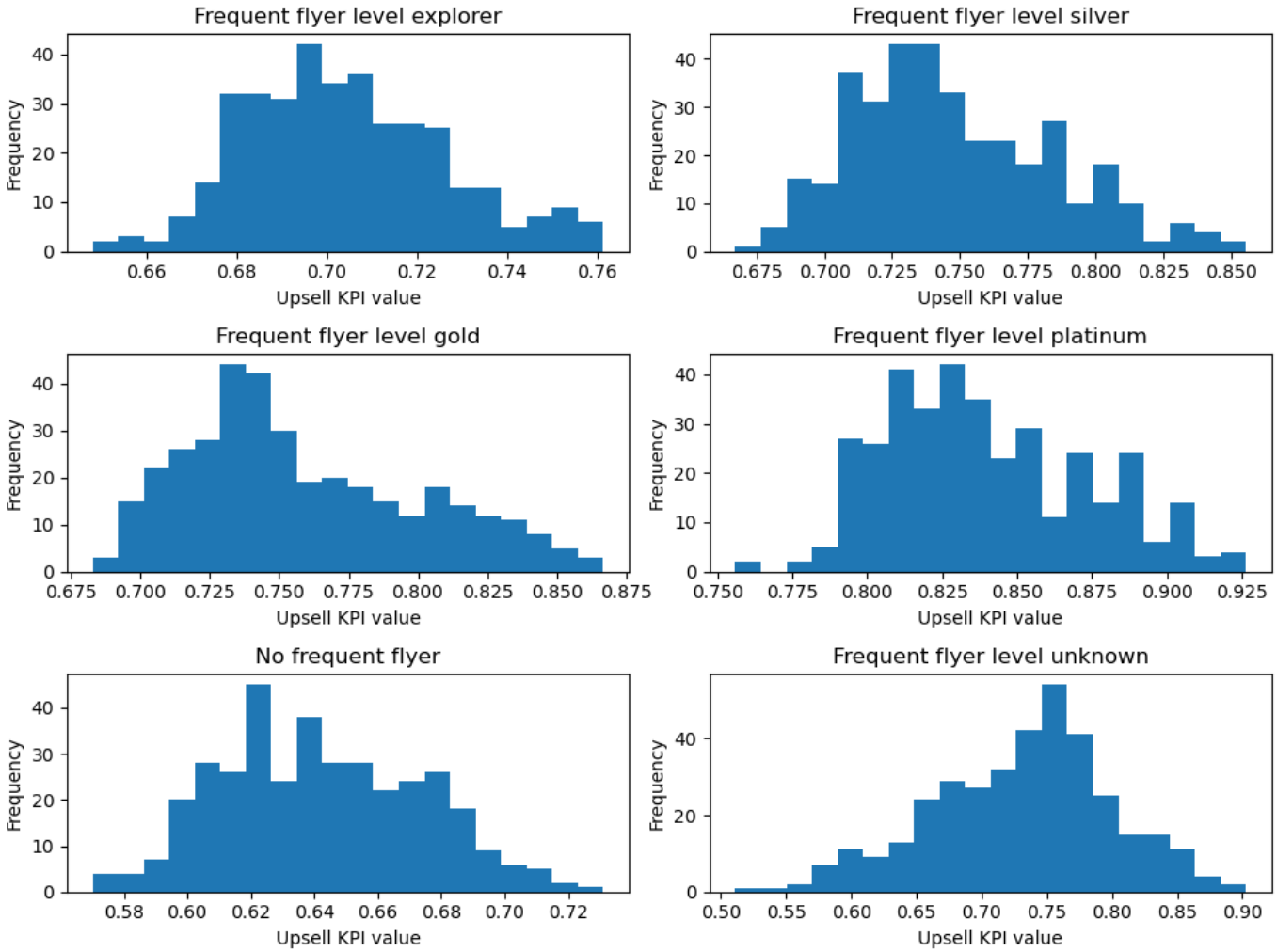


Figure 37: Histograms for frequent flyer level

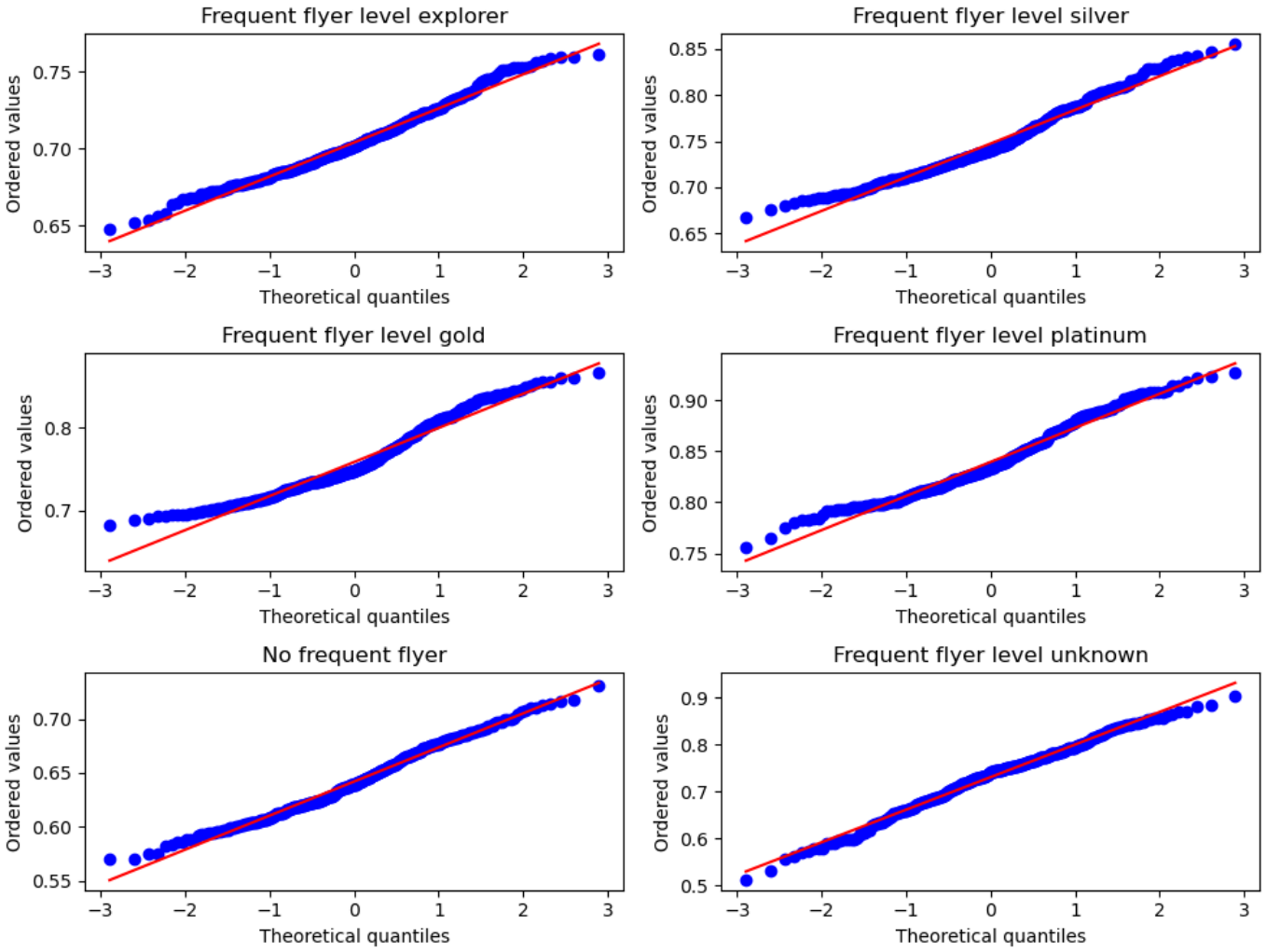


Figure 38: Q-Q plots for frequent flyer level

### A.1.9 Booking method

The booking method variable consists of five categories: direct online, indirect online, direct offline, indirect offline, and unknown. A box plot has been made for these categories and their corresponding upsell KPI. The box plot can be seen in figure 39, where we see that there is a big difference in the upsell KPI. Especially the indirect online category is different from the other categories.

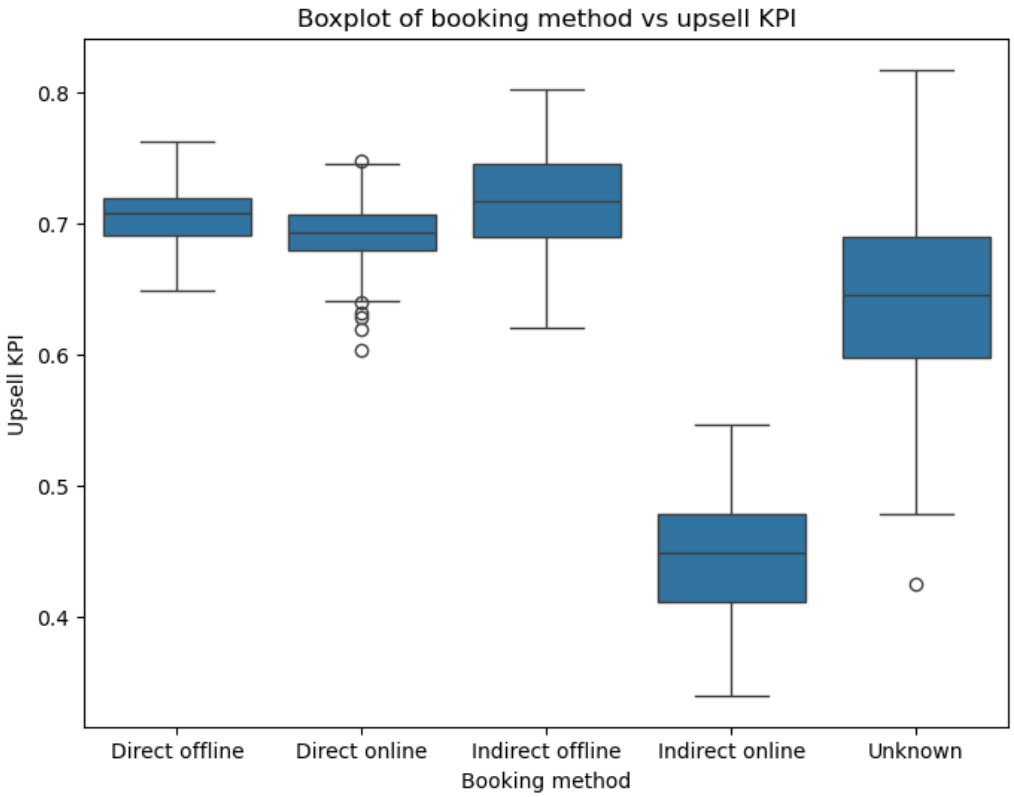


Figure 39: Box plot for booking method

First, we check for normality. For this purpose, histograms and Q-Q plots have been made. These can be found in figures 40 and 41. It can be seen that the data is indeed normally distributed. Next, a one-way ANOVA has been performed to check for a significant difference. A p-value below 0.05 has been found, and thus it is concluded that there exist significant differences between the different booking methods, and therefore this variable will be considered by the forecasting model.

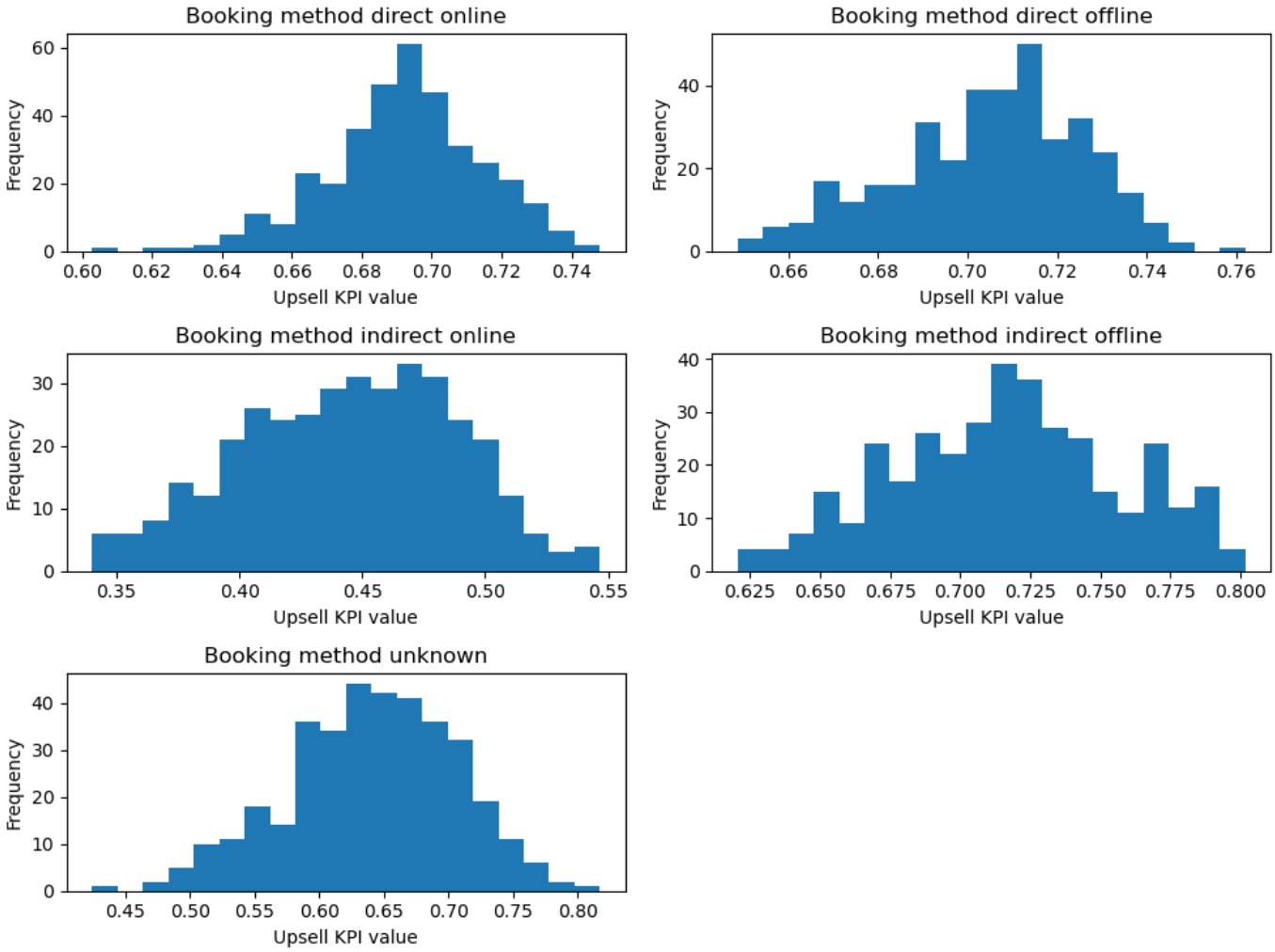


Figure 40: Histograms for booking method

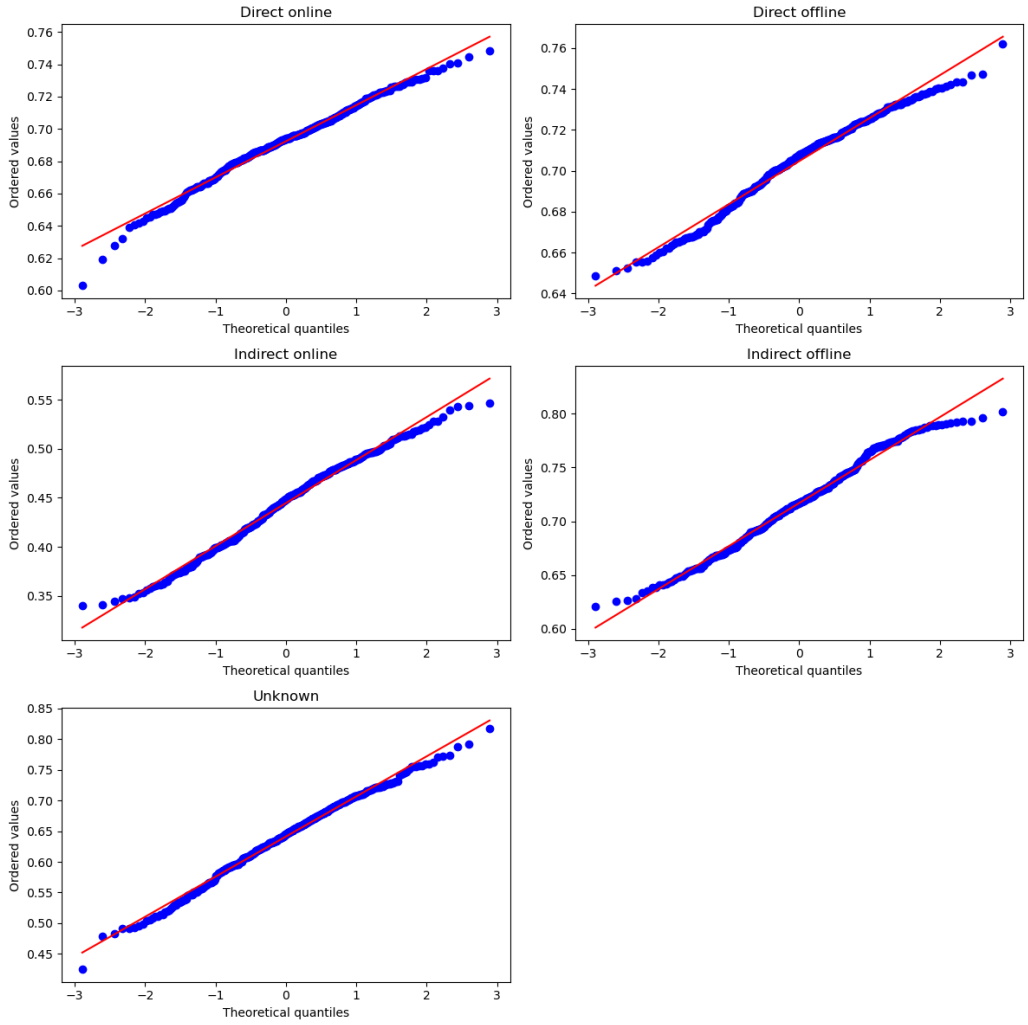


Figure 41: Q-Q plots for booking method

### A.1.10 Subclass

The subclass category consists of 27 categories, of which the 26 letters of the alphabet and the unknown category. In figure 42, a box plot of these categories versus their upsell KPI can be found. We can see quite some differences between the subclasses and their upsell KPI.

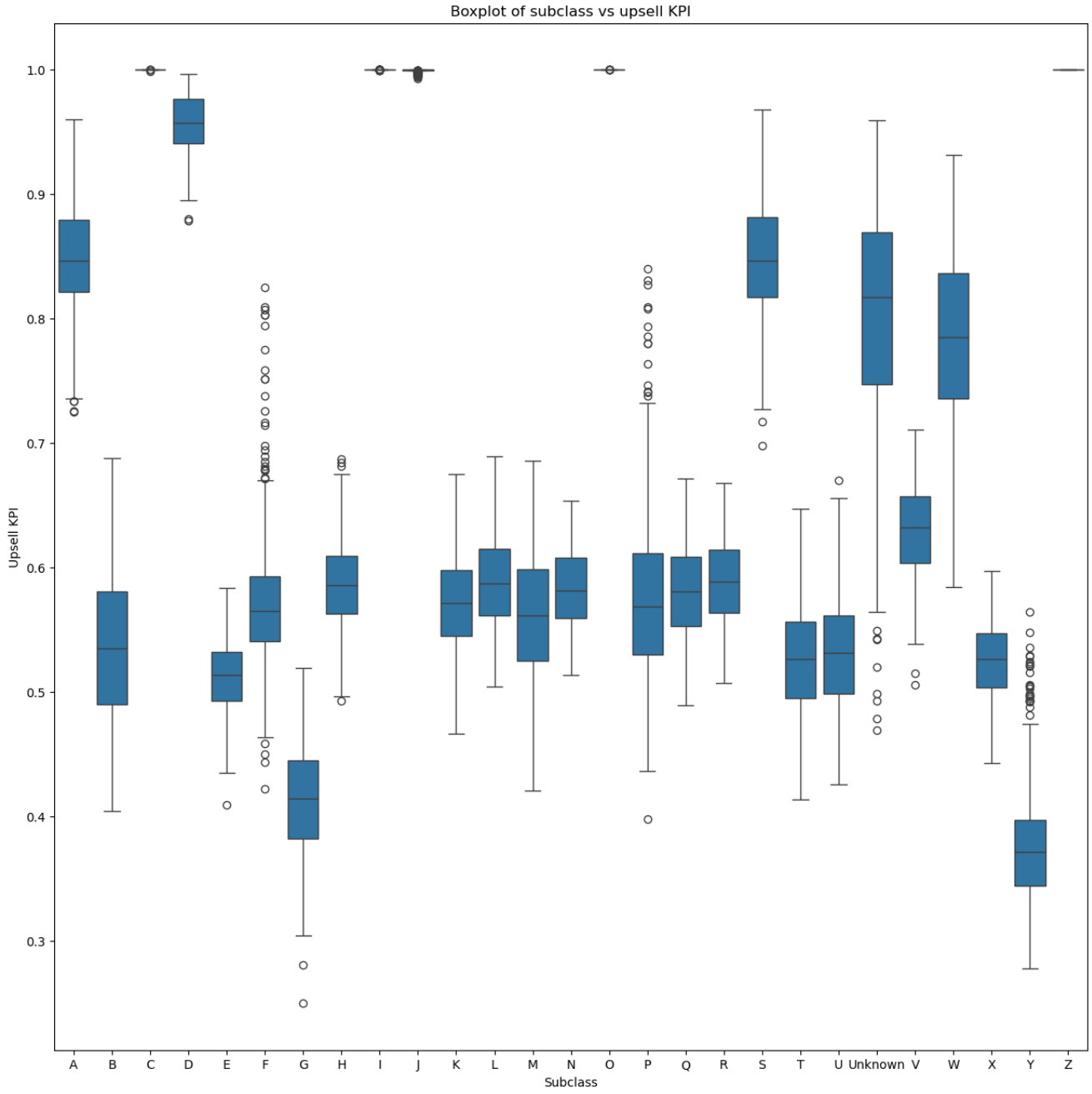


Figure 42: Box plot for subclass

Histograms and Q-Q plots are made for the different subclasses to verify normality. We can see

them in figures 43 and 44 and they show that most subclasses are normally distributed, but not all of them. Therefore, the Kruskal-Wallis test has been used to check for significant differences. A p-value below 0.05 has been found, and thus there are significant differences between the subclasses.

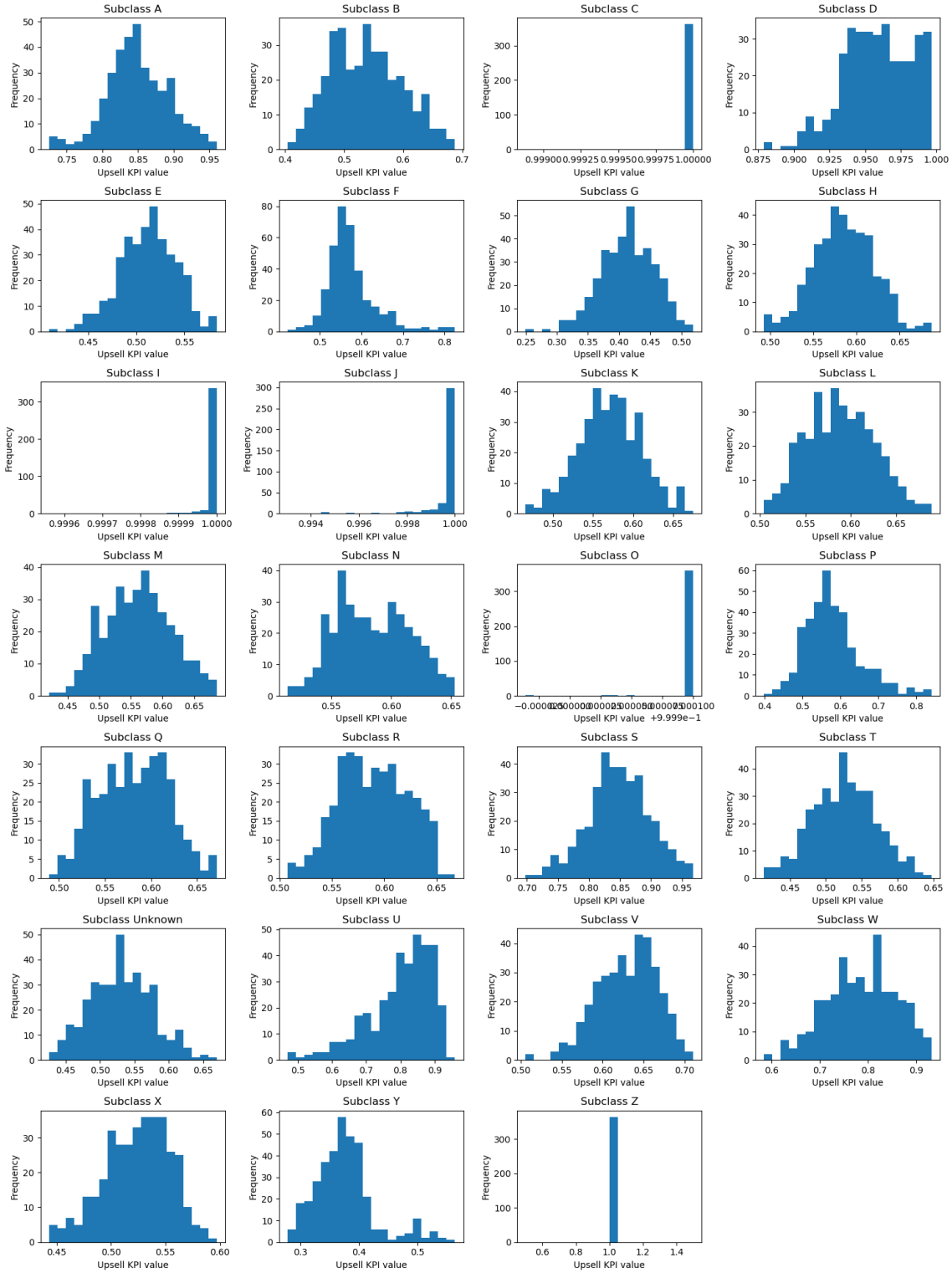


Figure 43: Histograms for subclasses

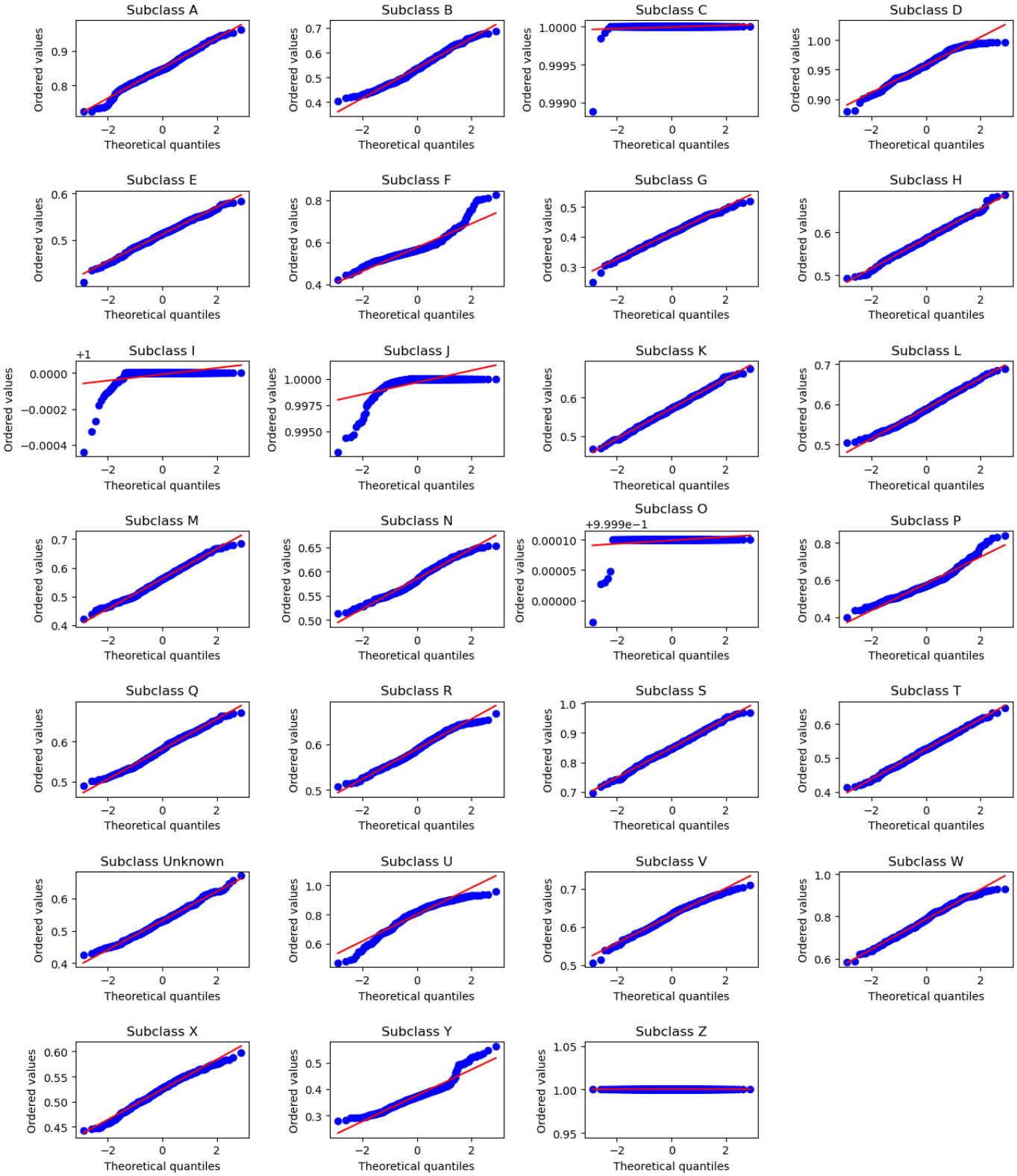


Figure 44: Q-Q plots for subclass

### A.1.11 Farebase season

The farebase season variable consists of eight categories. A box plot of these categories and their corresponding upsell KPIs can be found in figure 45. We see that there could be differences between the different farebase seasons.

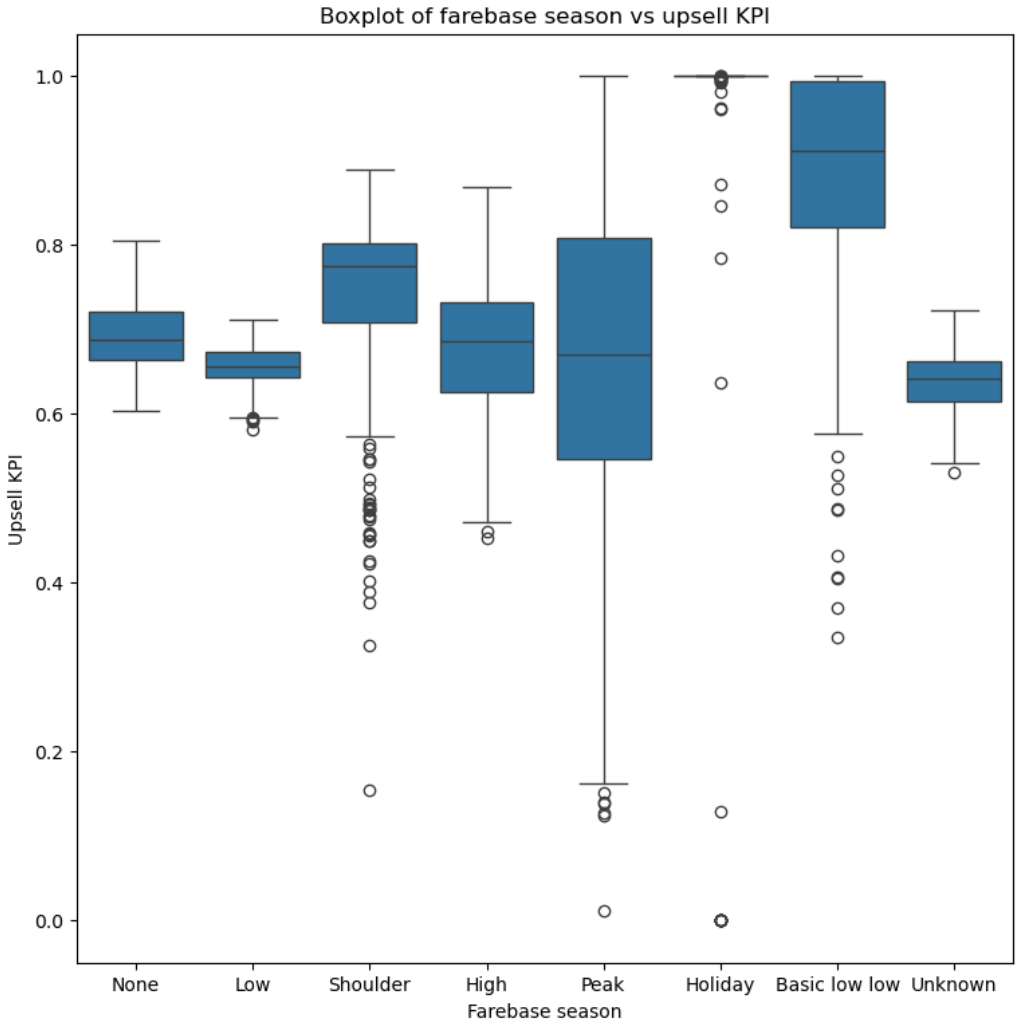


Figure 45: Box plot for farebase season

To verify whether there are significant differences between the farebase seasons, we first check for normality. To do this, the histograms in figures 46 and 47 have been made. They show that not

all seasons are normally distributed, and therefore the Kruskal-Wallis test will be performed. A p-value of 8.73e-263 has been found and therefore we conclude that there exist significant differences between the farebase seasons.

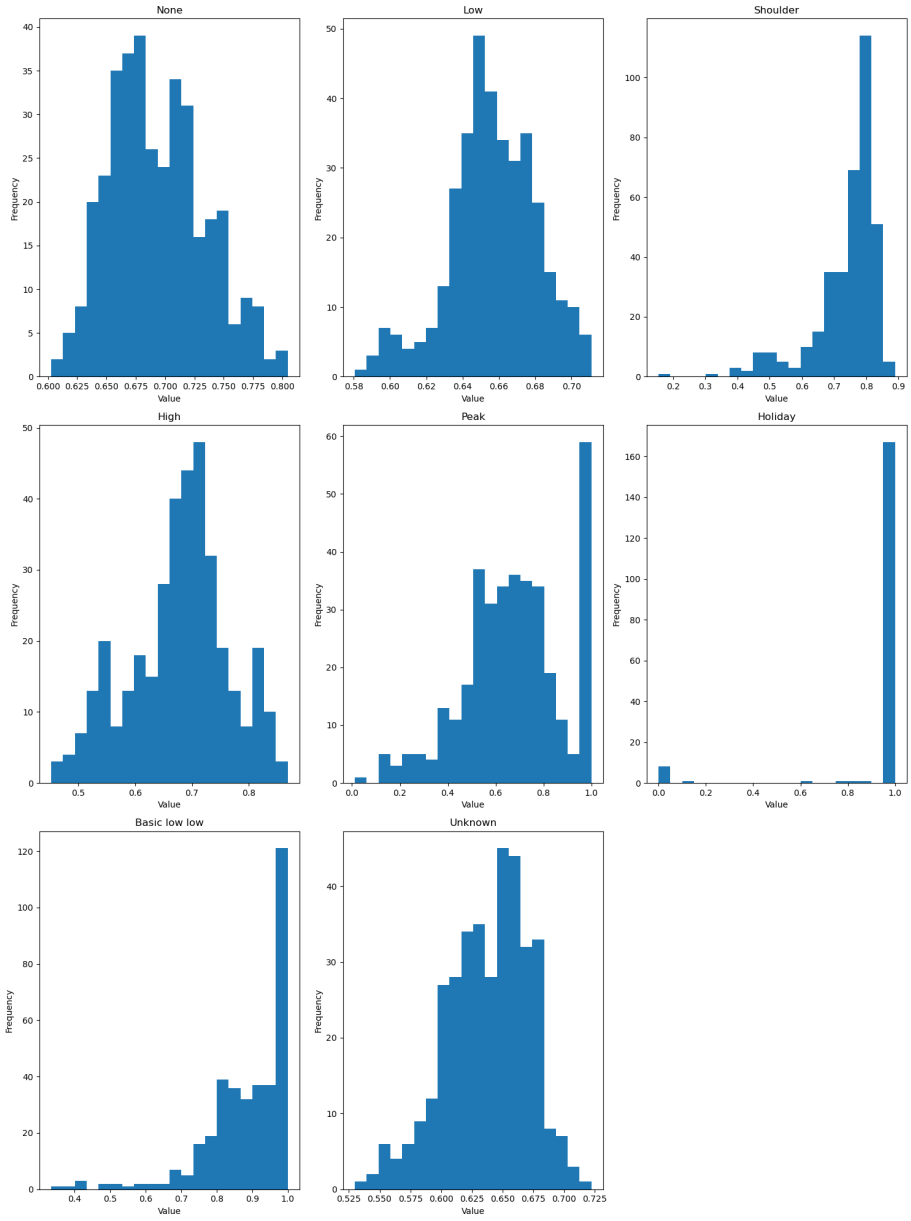


Figure 46: Histograms for farebase season

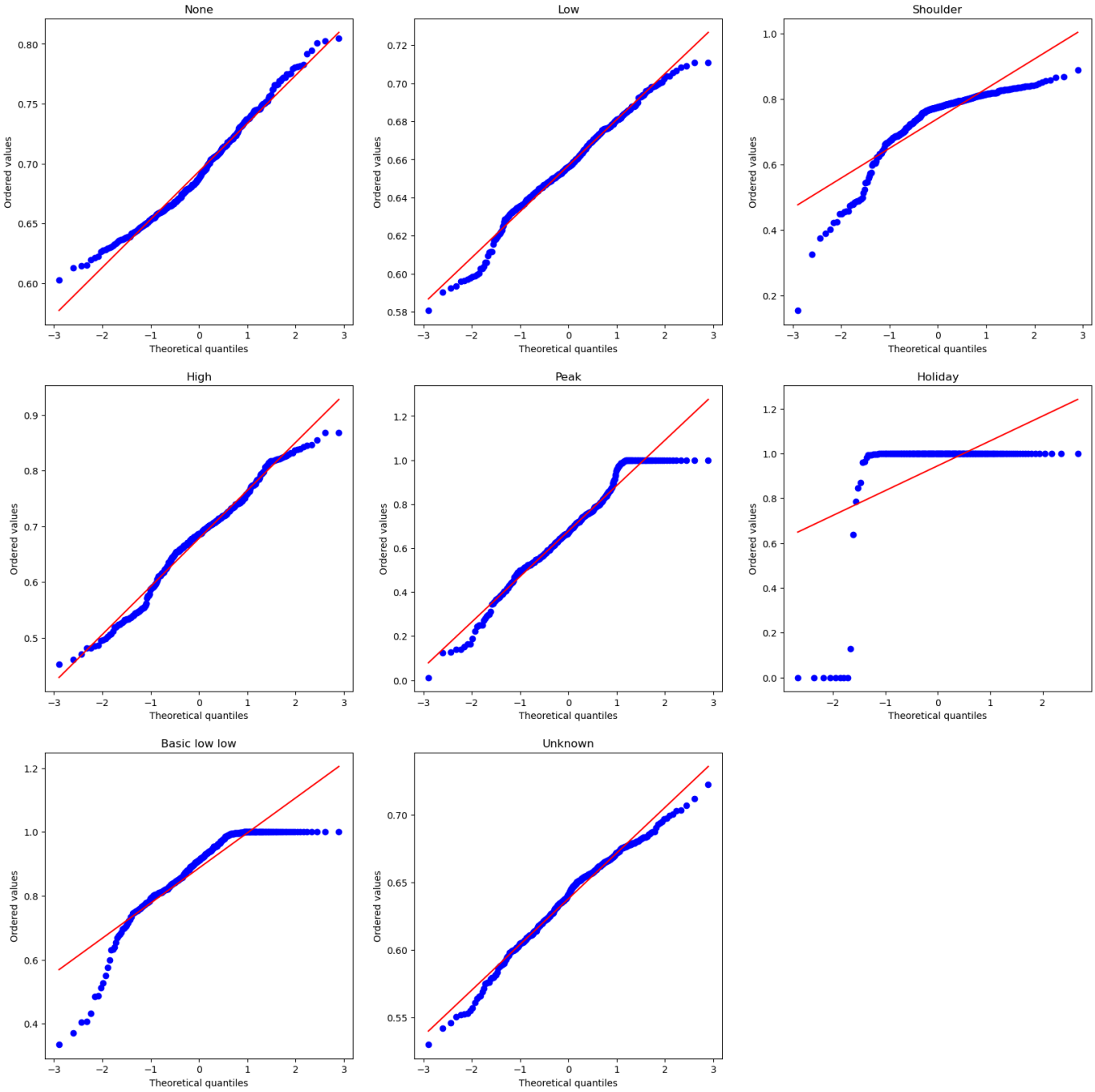


Figure 47: Q-Q plots for farebase season

### A.1.12 Traffic type

The traffic type variable consists of eight different types. These have been plotted against their respective upsell KPIs in figure 48. We can see quite some differences between the traffic types.

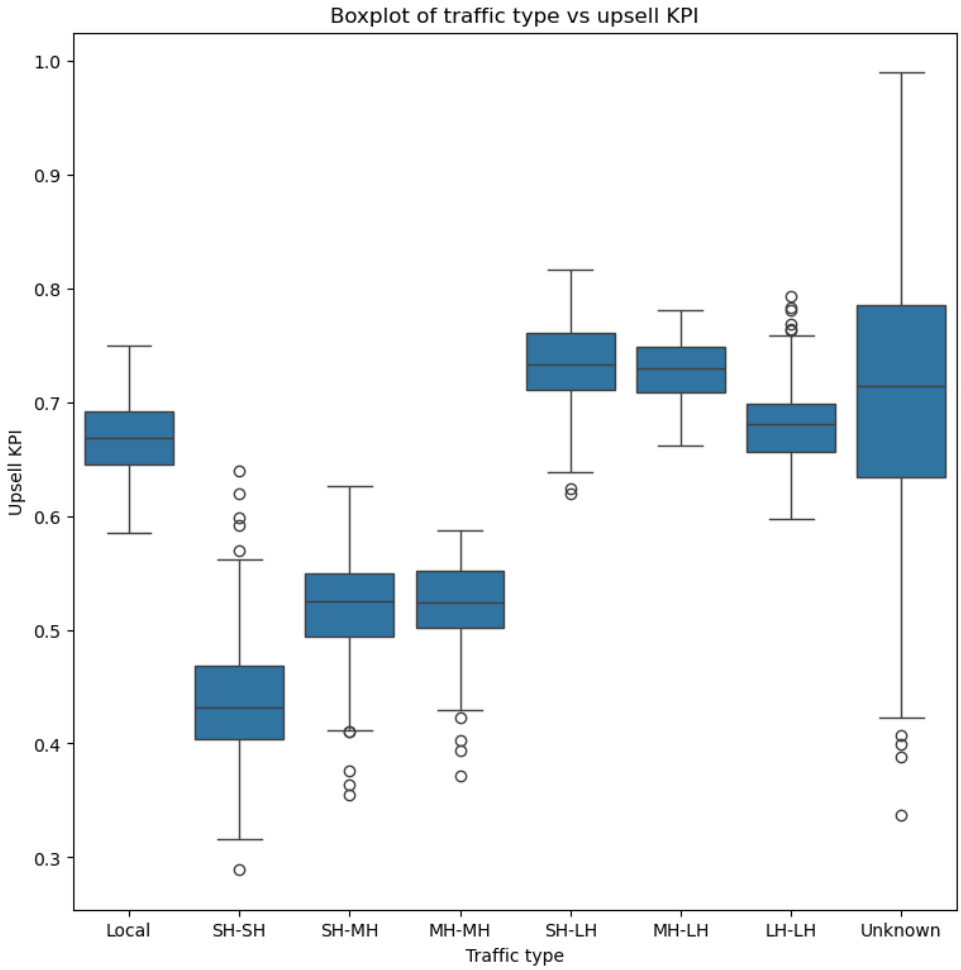


Figure 48: Box plot for traffic type

To check for normality, the histograms and Q-Q plots for these different traffic types have been plotted. We can see that all types are normally distributed, and therefore a one-way ANOVA test has been performed. The test gives a p-value below 0.05 and, therefore, it is concluded that a

significant difference exists between the traffic types.

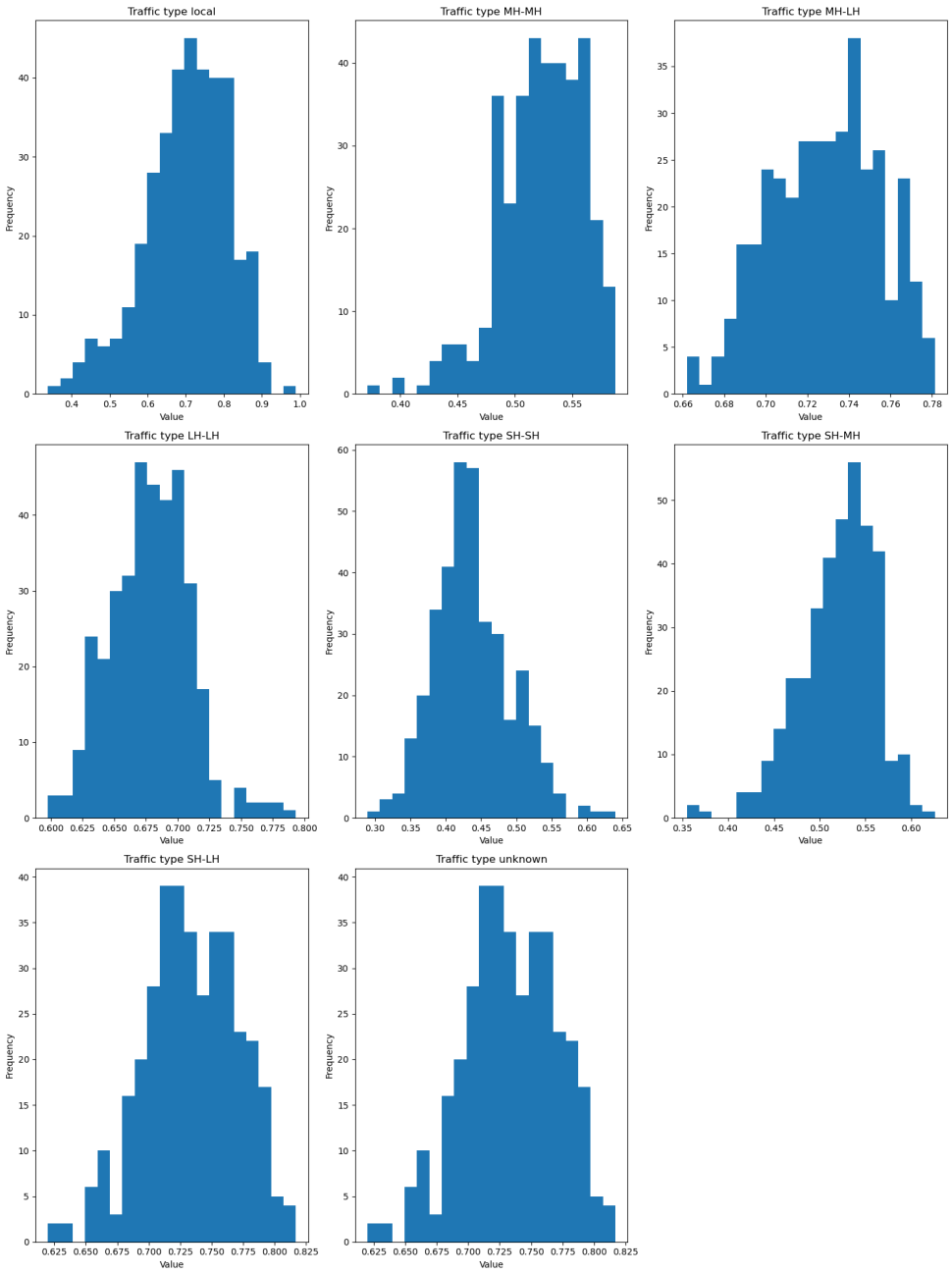


Figure 49: Histograms for traffic type

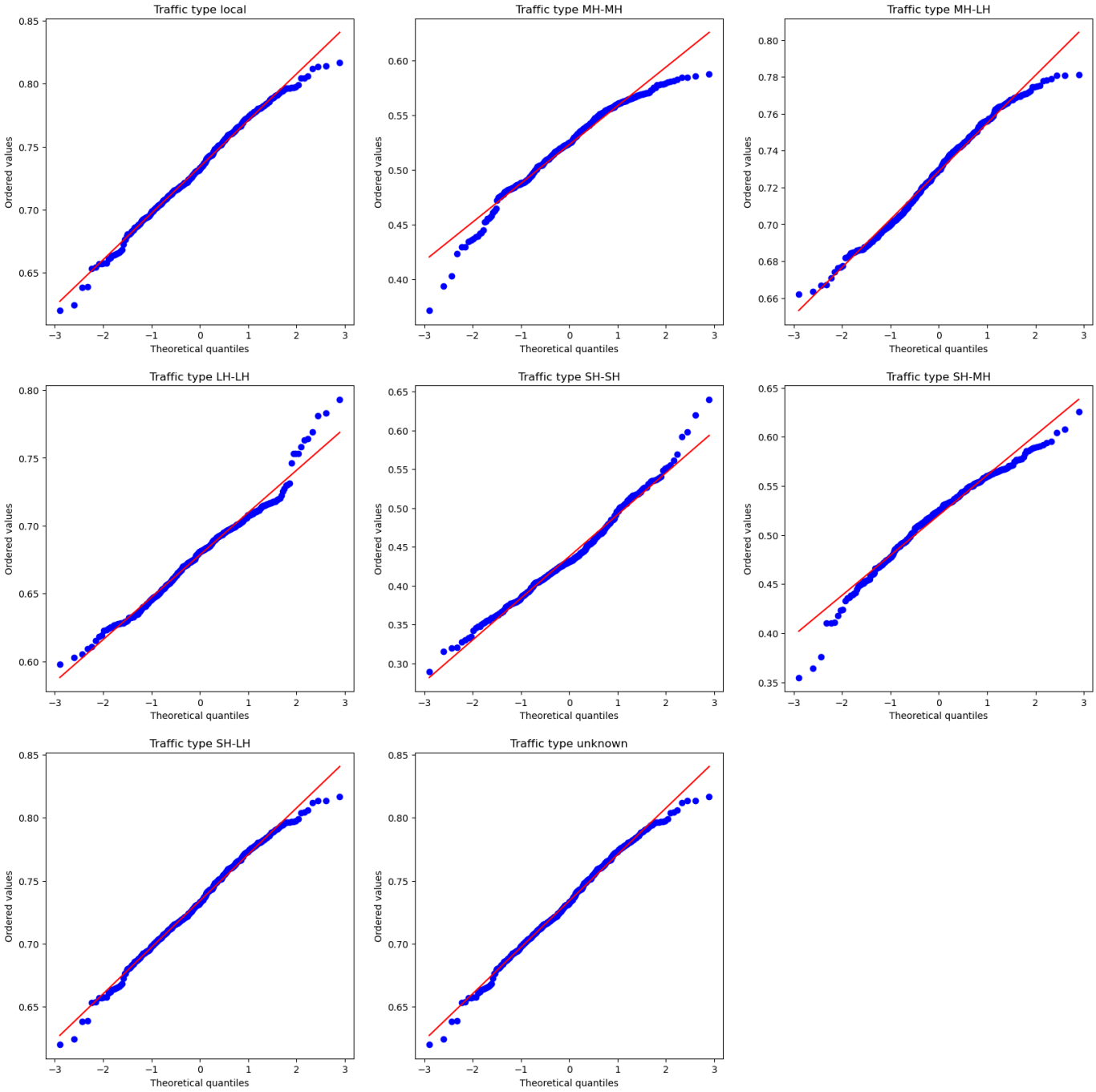


Figure 50: Q-Q plots for traffic type

## A.2 Results

### A.2.1 Feature importance forecast upsell KPI A without passenger information

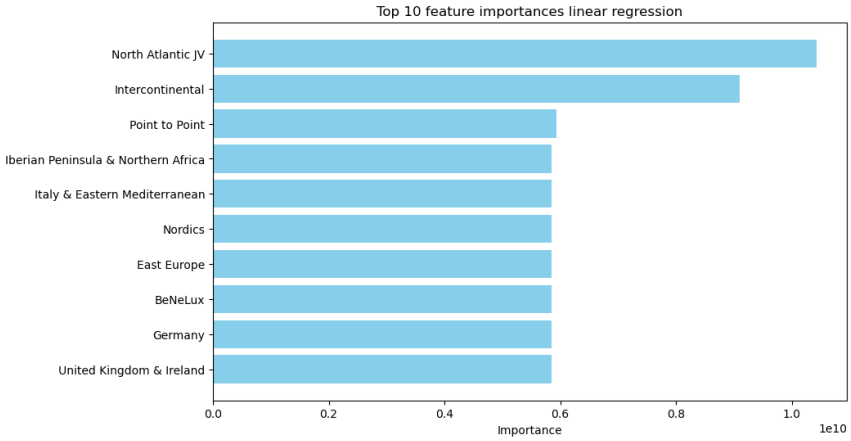


Figure 51: Feature importance for linear regression upsell KPI A

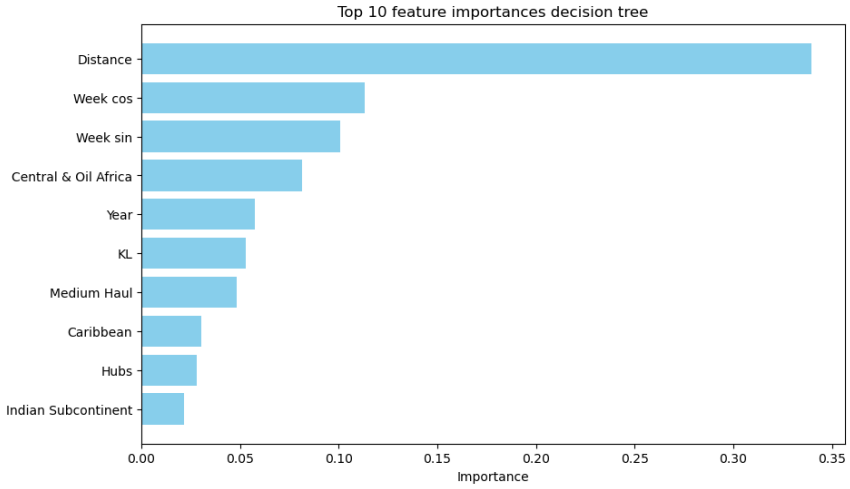


Figure 52: Feature importance for decision tree upsell KPI A

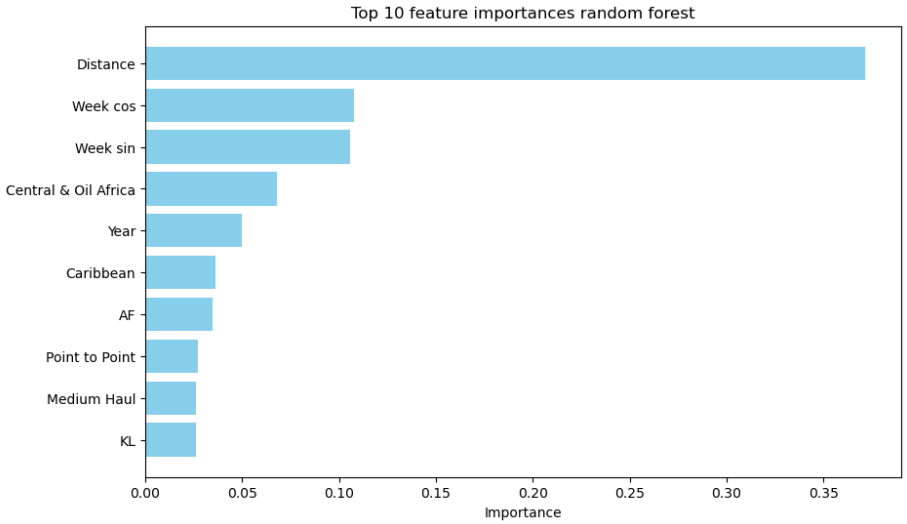


Figure 53: Feature importance for random forest upsell KPI A

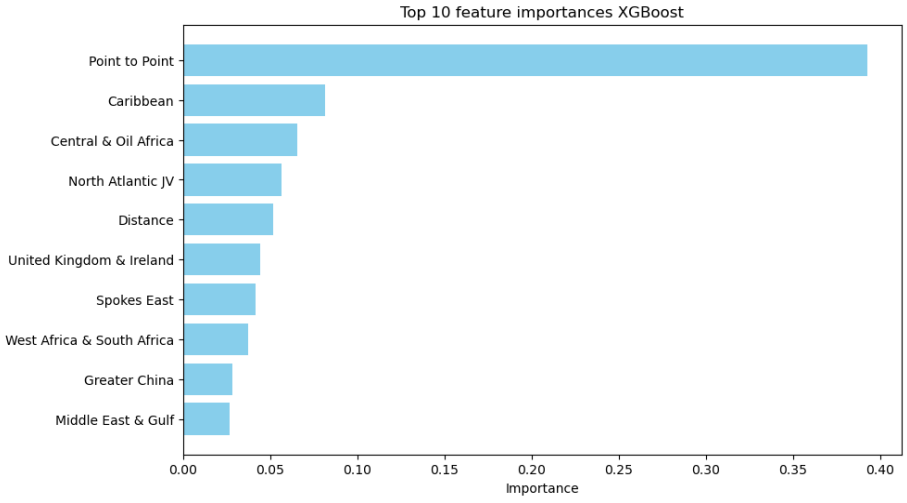


Figure 54: Feature importance for XGBoost upsell KPI A

### A.2.2 Feature importance forecast upsell KPI B without passenger information

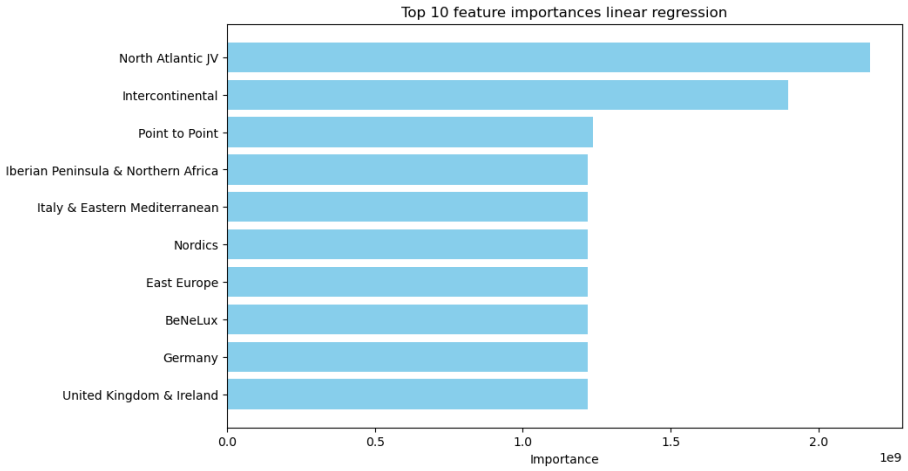


Figure 55: Feature importance for linear regression upsell KPI B

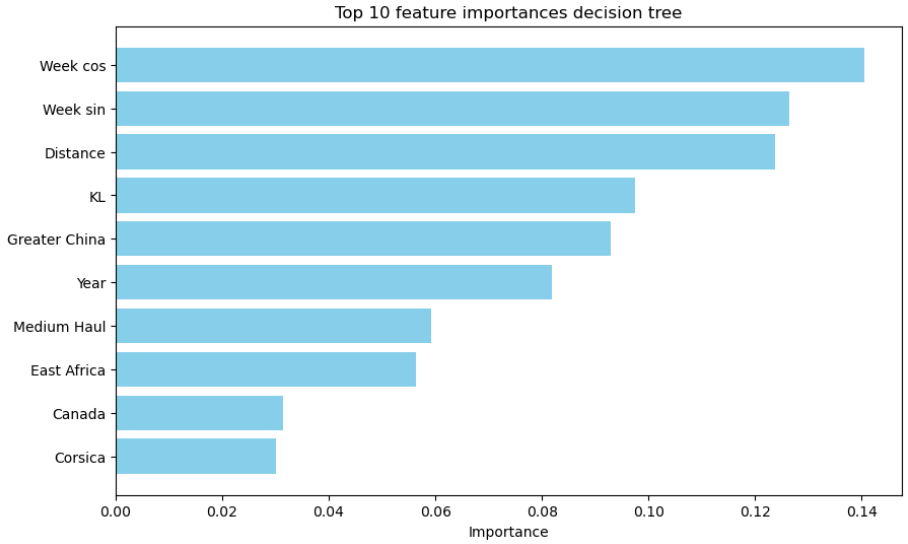


Figure 56: Feature importance for decision tree upsell KPI B

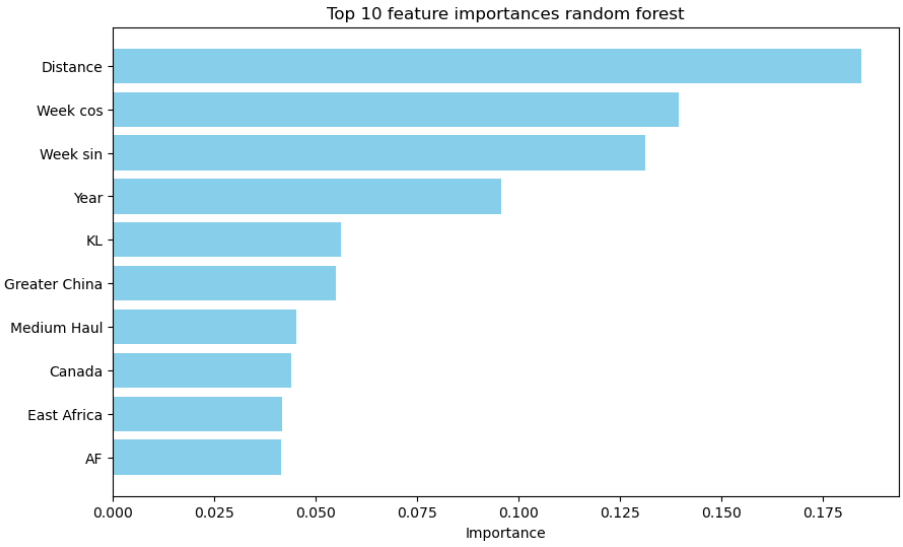


Figure 57: Feature importance for random forest upsell KPI B

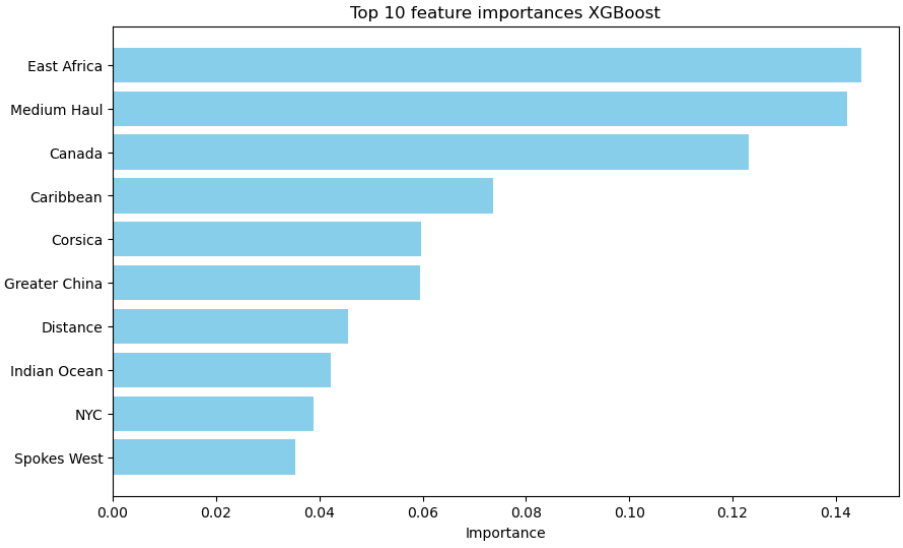


Figure 58: Feature importance for XGBoost upsell KPI B

### A.2.3 Feature importance forecast upsell KPI C without passenger information

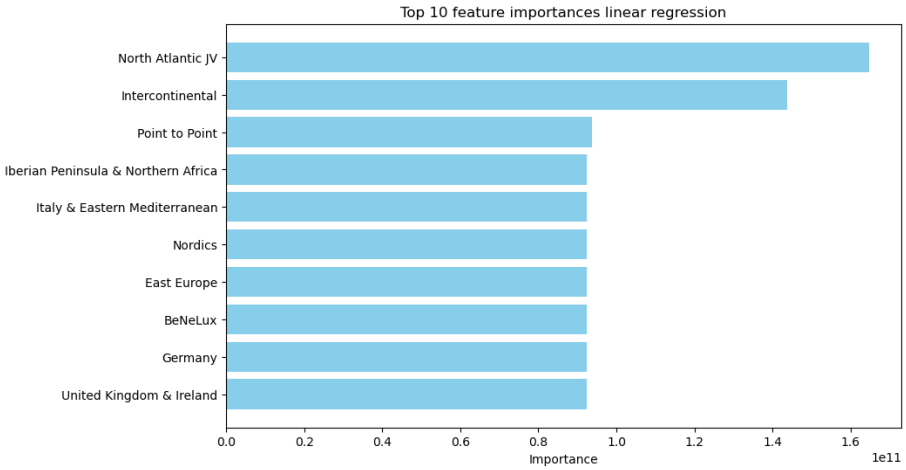


Figure 59: Feature importance for linear regression upsell KPI C

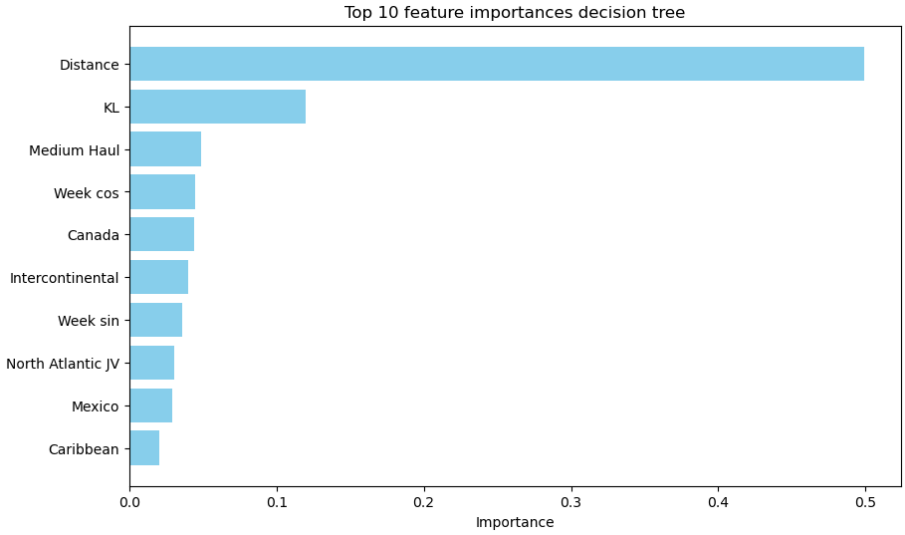


Figure 60: Feature importance for decision tree upsell KPI C

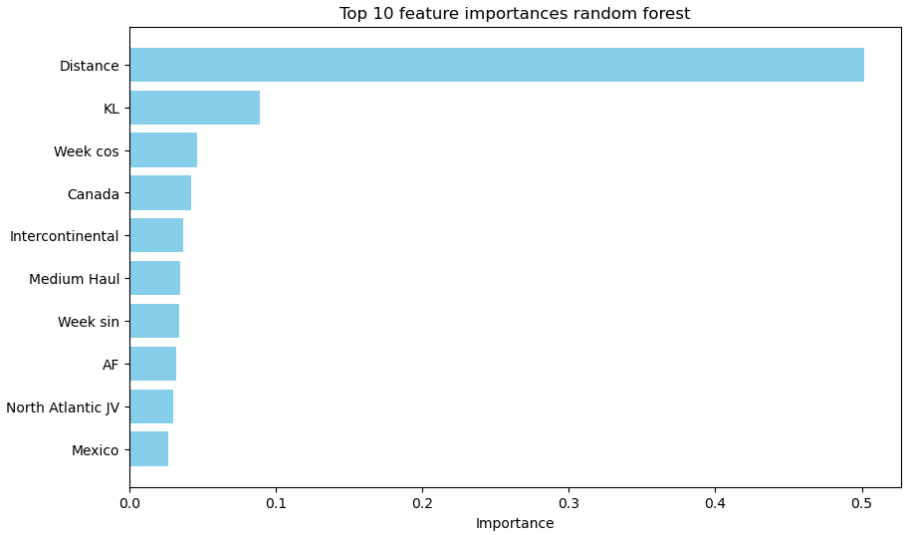


Figure 61: Feature importance for random forest upsell KPI C

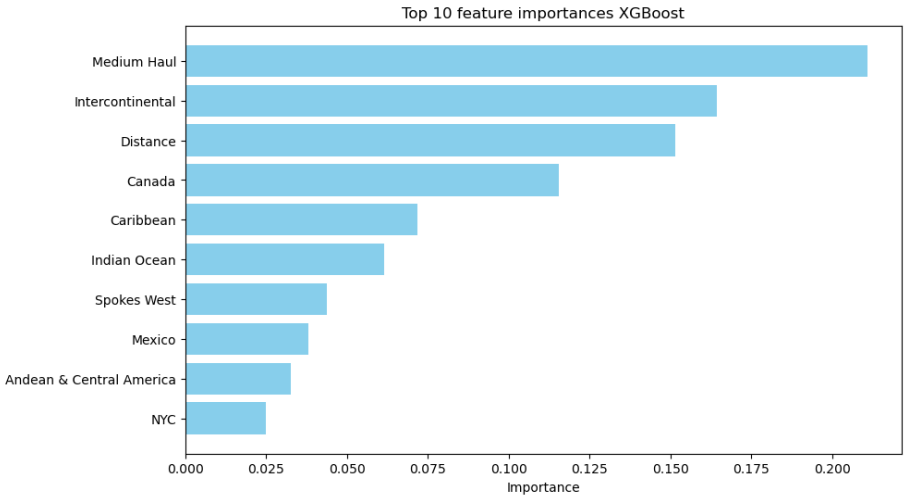


Figure 62: Feature importance for XGBoost upsell KPI C

### A.2.4 Feature importance forecast total upsell KPI without passenger information

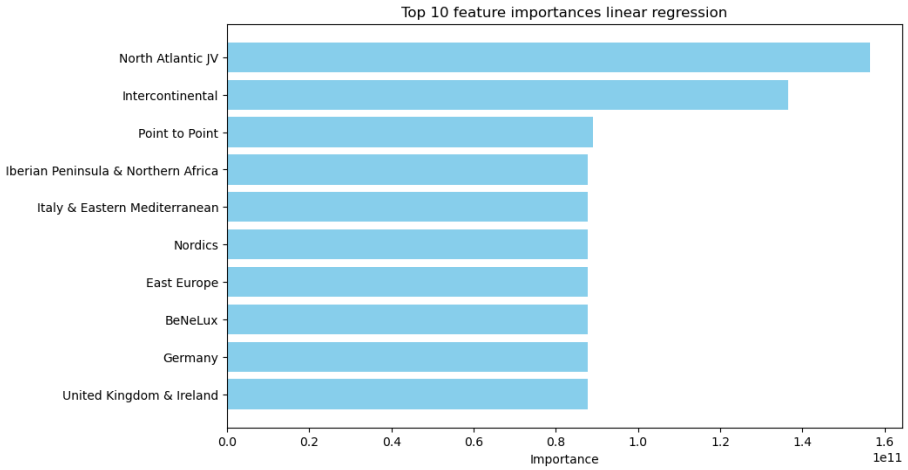


Figure 63: Feature importance for linear regression total upsell KPI

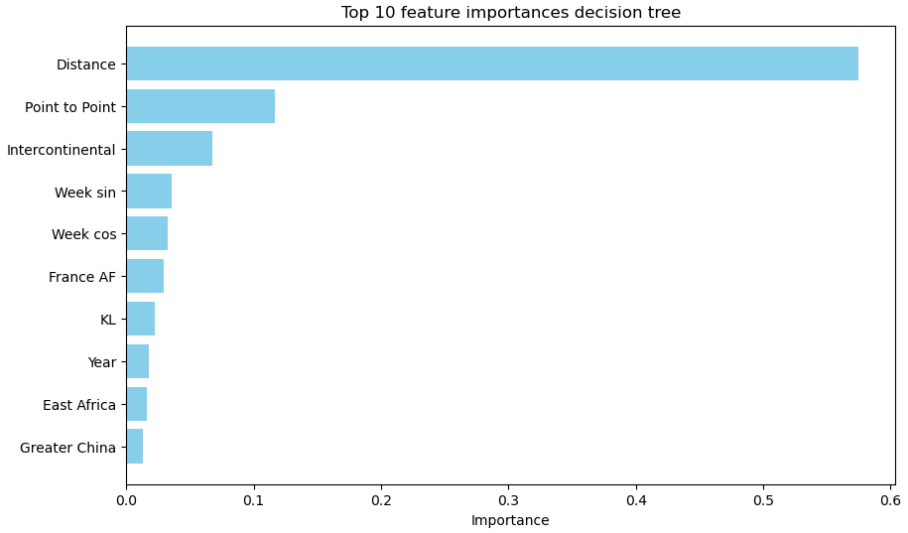


Figure 64: Feature importance for decision tree total upsell KPI

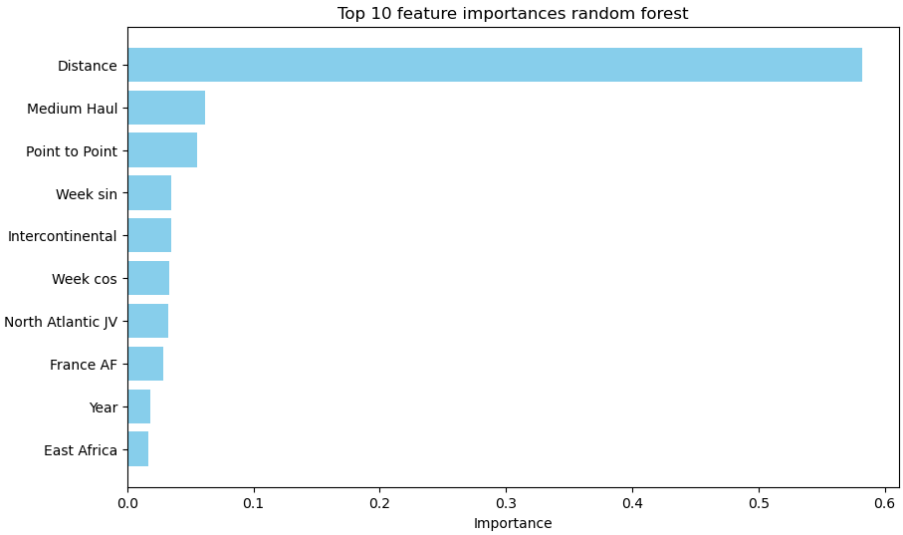


Figure 65: Feature importance for random forest total upsell KPI

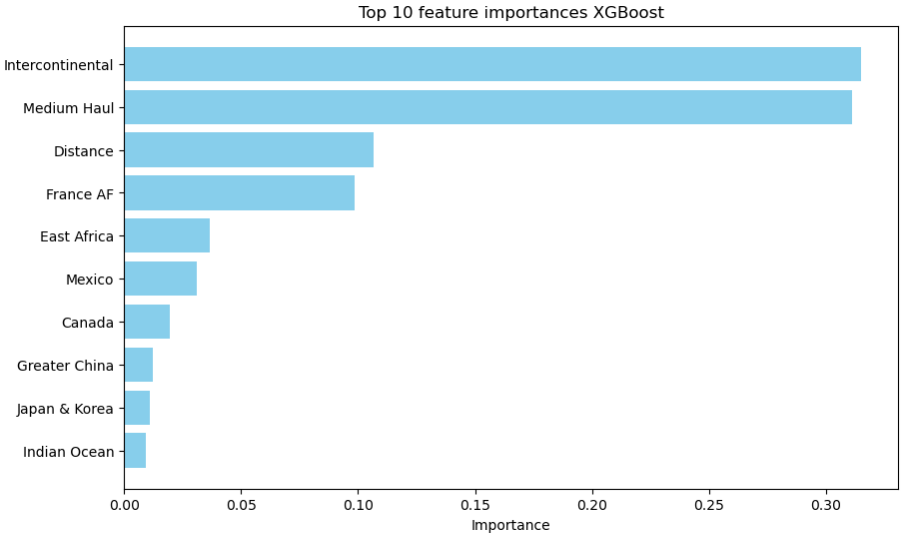


Figure 66: Feature importance for XGBoost total upsell KPI