The prediction of show ups at the doctor

Vrije Universiteit Amsterdam

Suzanne van der Velden (2565156)

Mentor: Sandjai Bhulai

Abstract

In this paper, a data analysis is conducted to predict the show ups at the doctor's office. Several features could have a potential influence on the show up. Based on the research outcomes of this paper, a model is created to predict the show up of patients. The model did not lead to solid predictions because of a potential imbalance dataset. Therefore, different methods to deal with imbalanced data are used. However, the issue could not be solved which resulted in an alternative suggestion: the Bailey-Welch rule.

Contents

1. Introduction

It seems like a straight forward process. A patient calls the doctor for a medical problem, the appointment is scheduled and the doctor's visit is executed. However, it happens quite often that patients do not show up at the appointment. This means the doctor will have an empty spot in the agenda and the patient pays a fine which can easily run up to 50 euros (Huisarten Praktijk Presikhaaf, 2017). There are no winners in this case, but who to blame?

This paper shows the findings of the research conducted on the dataset "Medical appointment No-Shows". This dataset can be found on the Kaggle website (Hoppen, 2017). The purpose of this research is to predict if a patient will show up at his medical appointment or not and what to do if the show up is not predictable. The dataset contains 300,000 medical appointments with each 15 features, of which one is the target feature Status (Show ups). The target feature consists of "No-Shows" for 30.24% and 69.76% are "Show-ups". The other features might have influence on the actual show up rate. In order to analyse the impact of these features, several experiments are conducted.

Firstly, the dataset is explored. Each feature is discussed individually with exception of the seven medical characteristics included in the dataset. Then the most promising features, regarding the prediction of show ups, are chosen. Combinations of these features are used in the prediction model. Finally, conclusions are made about the test results of the prediction model.

2. Data analysis

A data analysis is performed first. In this chapter, an overview of the features is given. The first feature is the patient's age. Six out of the 300,000 age values were negative. Because it is not possible for a patient to have an age below zero, they are considered as errors in the dataset. The decision is made to exclude these patients from the data. From now on, the dataset consists of 299,994 medical appointments. In Figure 1, two plots are shown. The one on the left shows the distribution of the age when the patient shows up at his appointment. The plot on the right shows the distribution of the age when a patient does not show up at his appointment. As can be seen, the plot with the age of the patient when he shows up is a little more skewed to the right. This can mean that the probability of an older patient showing up is bigger than the probability of a younger patient showing up.

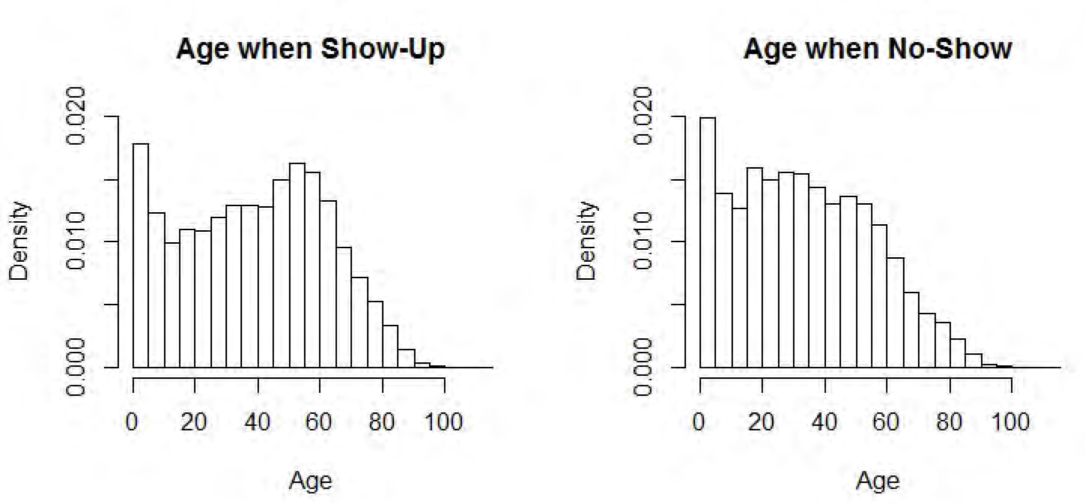

***Figure 1**: A density graph of the age of a patient with respect to the Show up or No Show of a patient.*

The numbers from Table 1 also show that the average age of the patients who show up is higher than the average age of the patients who do not show up at their appointment. This also holds for the median and the standard deviation.

Age	Show-Up	No-Show	Total population
Mean	39.308	34.351	37.809
Median	41	33	38
Std. dev.	23.082	21.775	22.809

***Table 1**: The mean, median, and standard deviation of the feature Age.*

To test if the distribution of the age of a patient who shows up at his appointment is significantly different from the distribution of the age of a patient who does not show up, the Kolmogorov-Smirnov test (ks test) is performed. The ks test checks if two data samples belong to the same distribution. This is done by examining the cumulative distribution function (cdf) of the two data samples. The ks test checks if the cdf of the data samples are equal or not. In Figure 2 the cdf of the age of both the show-ups and no-shows is given. As can be seen, the cdf of the patients who show up lies below the cdf of the patients who do not show up at their appointment.

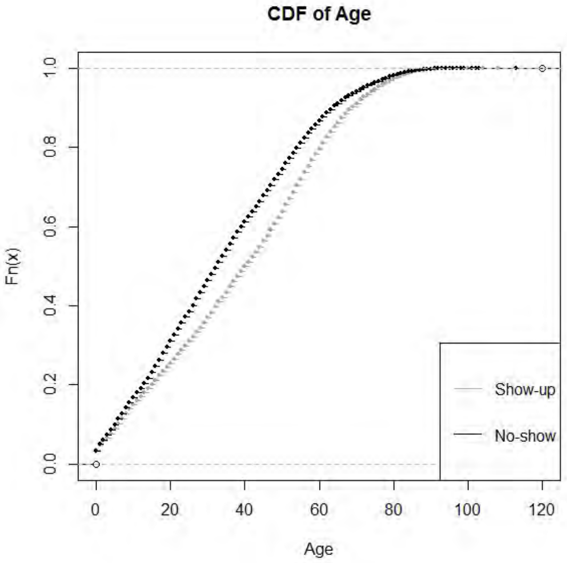

Figure 2*: The cumulative distribution function of the age, grouped by the status of a patient.*

The null-hypothesis for the ks test is defined as: the cdf of the age of patients who show up is equal to the cdf of the age of patients who do not show up at their appointment. The alternative hypothesis is: the cdf of the age of patients who show up is not equal to the cdf of the age of patients who do not show up at their appointment. The p-value of the ks test with this null and alternative hypothesis is <2.2e-16. This means that the cdf of the age of patients who show up at their appointment is significantly different from the cdf of the age of patients who do not show up at their appointment.

These findings can mean that the age of a patient may be a good feature to predict if the patient will show up at his appointment or not.

The second feature is the gender of a patient. Most of the appointments (66.83%) are made by females. From these appointments, 70.13% show up. This means that 33.17% of the appointments are for males. A male shows up in 69% of the cases. This is slightly less than the percentage of the females who show up.

The third and fourth features are the date and time the appointment is made, and the date the appointment will take place. The dataset already contains the feature waiting time, which is the difference in time between the date the appointment is made and the date the appointment will take place. The time of the appointment made is chosen to find a potential new feature. In Figure 3, two plots are shown. The one on the left shows the distribution of the time the appointment is made when the patient shows up at his appointment. The plot on the right shows the distribution of the time the appointment is made when the patient does not show up at his appointment. The two plots are almost similar. The main difference is the peak at the beginning of the day for a patient who shows up.

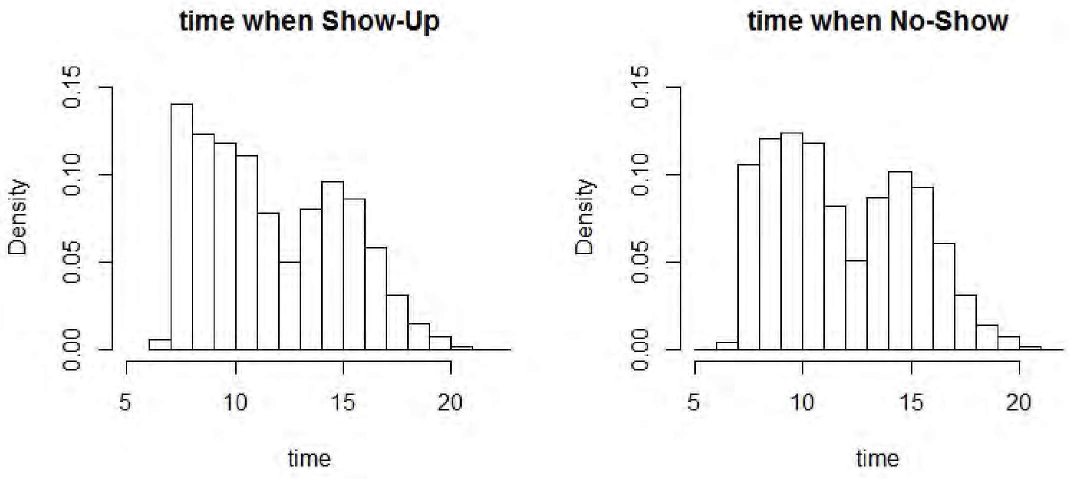

Figure 3: *A density graph of the time a patient made an appointment with respect to the Show up or No Show of a patient.*

The numbers from Table 2 show that there is not a big difference between the average of the time when the patient shows up or not. This also holds for the median and standard deviation. Seeing these results, the feature of the time when the appointment is made will likely not be used to predict the show-ups.

Time	Show-Up	No-Show	Total population
Mean	11.702	11.905	11.764
Median	11.033	11.317	11.117
Std. dev.	3.253	3.176	3.231

Table 2: *The mean, median, and standard deviation of the feature "time the appointment is made".*

The fifth feature is the day of the week the appointment is taking place. The distribution of the no shows and show ups can be found in Table 3. As can be seen, the percentage no shows on Saturday and Sunday are different from the percentage no shows on weekdays. This can be explained by the fact that Saturdays and Sundays occur less often in the dataset than weekdays. Medical appointments are on Saturday or Sunday in less than 1% of the cases. Furthermore, the percentage of show ups on Mondays is 3.22% lower than on Tuesdays. This difference can be useful to predict if a patient will show up at his appointment.

Day of the week	% Show up	% No show	% of total dataset
Monday	67.87	32.13	19.77
Tuesday	71.09	28.91	20.93
Wednesday	70.20	29.80	21.17
Thursday	70.38	29.62	20.09
Friday	69.21	30.79	17.59
Saturday	63.17	36.83	0.46
Sunday	83.33	16.67	0.002

Table 3*: The distribution of the no shows and show ups per day and the percentage of how often the day occurs in the dataset.*

The sixth feature is the 'status'. This is the target feature. The target feature consists for 30.24% out of "No-Shows" and 69.76% are "Show-ups". Feature seven till thirteen are all binary features except for the tenth feature. These features encode if the patient has diabetes, alcoholism, hypertension, a handicap, smokes, a scholarship, or tuberculosis. The probability of a patient showing up at his appointment against one of the features is given in Figure 4. The most striking fact is that the probability of showing up is 12.10% bigger when the patient has handicap 4 instead of handicap 0, where 0 is no handicap and 4 is a severe handicap. This can be explained by the fact that if a patient has a severe handicap, he probably has to be brought to the medical appointment by someone else. The show-up will then depend on more than one person.

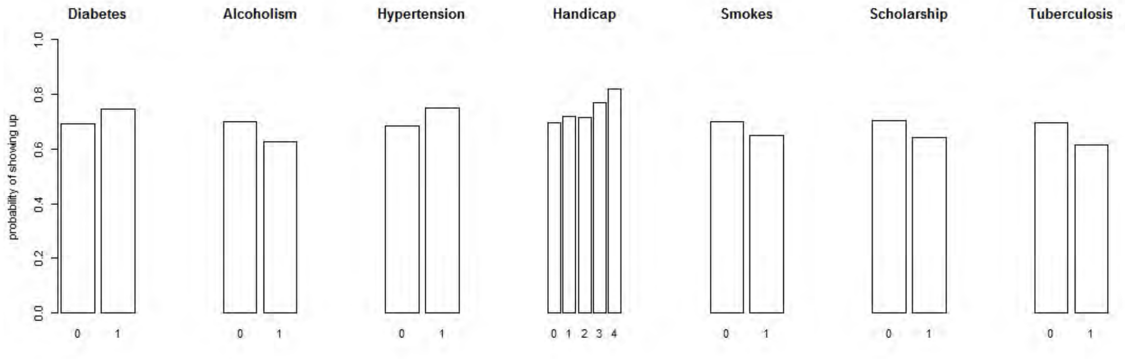

Figure 4: *The probability for a patient to show up at their medical appointment with respect to several features.*

The fourteenth feature is the sms reminder. Some patients receive a sms reminder before their medical appointment. This feature consists of '0', '1', and '2'. This is the number of reminders the patient has received. The distribution of this feature is shown in Table 4. As can be seen, the percentage show ups does not deviate much from each other in the different groups of the feature sms reminder. This can mean that the feature 'sms reminder' may not be suited to predict if a patient will show up.

Sms reminder	% Show up	% No show	% of total dataset
0	69.73	30.27	42.85
1	69.79	30.21	56.88
2	66.46	33.54	0.27

***Table 4**: The distribution of the no shows and show ups for the different values of the feature sms reminder.*

The fifteenth and last feature is the waiting time. This feature measures the days between the day the appointment is made and the day the appointment will take place. The left boxplot in Figure 5 shows the distribution of the feature waiting time. The boxplot in the middle shows all waiting times below 100. This is done because of the many outliers. The boxplot on the right shows how the boxplot would look like if all waiting times higher than 43 will be ignored. In this case, 4.14% of the data will be ignored. If a waiting time of 100 is the maximum, then only 0.28% will be ignored. Therefore, all waiting times higher than 100 are set to 100.

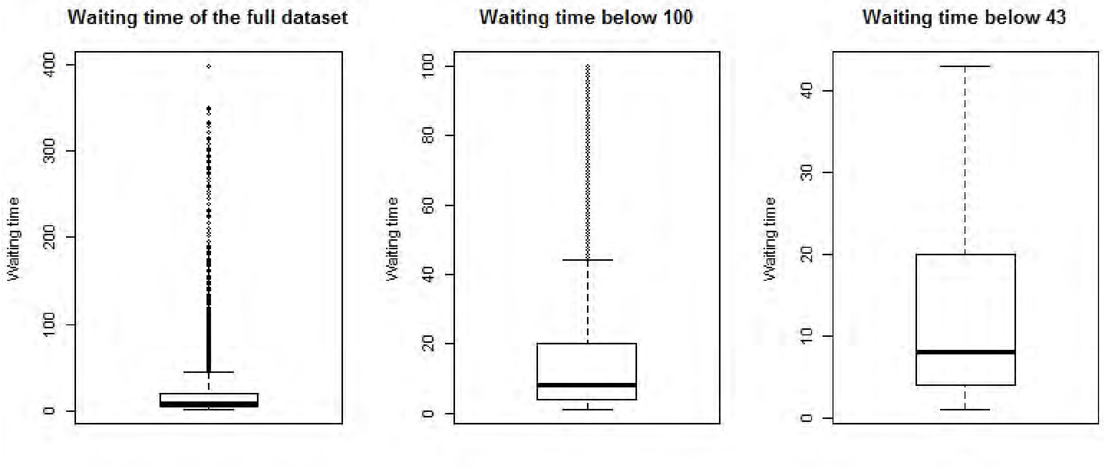

***Figure 5**: Three boxplots of the waiting time in days between the day the appointment is made and the day the appointment will take place. The one on the left shows the waiting time of the full dataset. The boxplot in the middle shows the waiting time without the ones bigger than 100. The boxplot on the right only shows the waiting time without the ones bigger than 43.*

In Figure 6, two plots are shown. The one on the left shows the distribution of the waiting time when the patient shows up at his appointment. The plot on the right shows the distribution of the waiting time when the patient does not show up at his appointment. The two plots do not differ much from each other, the only big difference is that in proportion to the right plot the area under the density of a waiting time from zero to five on the left plot is almost 2% lower. This can mean that if the waiting time is shorter than 5 days, a patient is more likely show up then when the waiting time is longer than 5 days.

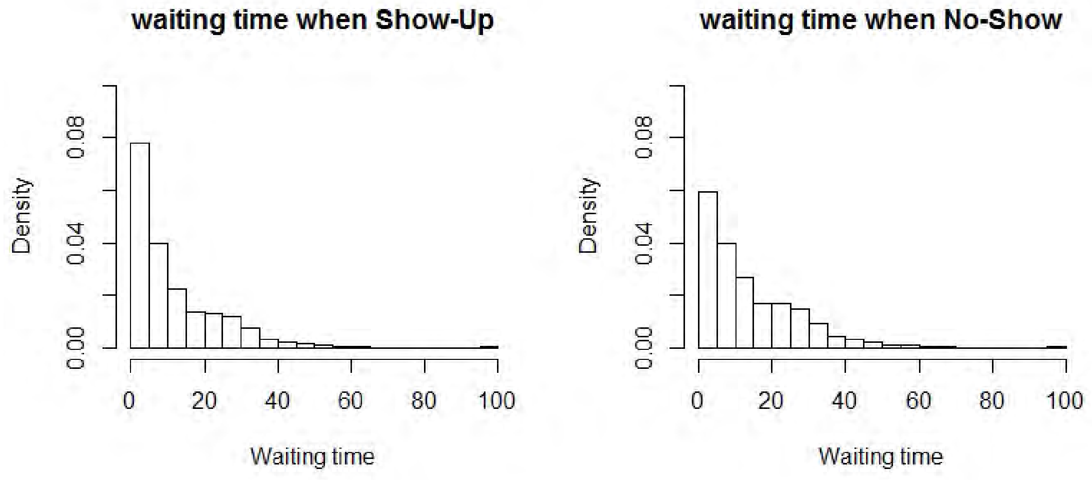

***Figure 6**: A density graph of the number of days between the appointment made and the appointment will take place with respect to the Show up or No Show of a patient.*

The mean waiting time for the patients who show up is more than two days shorter than for patients who do not show up at their appointment. This also holds for the median and the standard deviation. This can be seen in Table 5. This could mean that the number of days between the day that the appointment is made and the day the appointment will take place can be a good feature to predict if a patient will show up at his appointment or not.

Time	Show-Up	No-Show	Total population
Mean	13.06	15.33	13.74
Median	7	11	8
Std. dev.	14.47	15.10	14.70

***Table 5**: The mean, median, and standard deviation of the feature "waiting time".*

The ks test is also performed on the feature waiting time. As can be seen in Figure 7, the cdf of the waiting time of the patients who show up at their appointment lies above the cdf of the patients who do not show up at their appointment. To check if this is significant, the ks test is performed with the following null hypothesis: the cdf of the waiting time of patients who show up is equal to the cdf of the waiting time of patients who do not show up at their appointment. The alternative hypothesis is: the cdf of the waiting time of patients who show up is not equal to the cdf of the waiting time of patients who do not show up at their appointment. The p-value of the ks test with these null and alternative hypothesis is <2.2e-16. This means that the cdf of the waiting time of patients who show up at their appointment is significantly different from the cdf of the waiting time of patients who do not show up at their appointment.

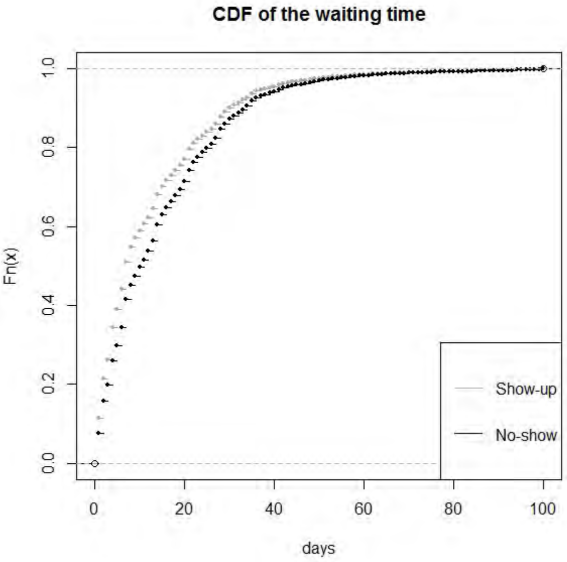

***Figure** 7: The cumulative distribution function of the waiting time, grouped by the status of a patient.*

3. Predicting

After the data analysis is performed, it is time to start predicting the “show-ups”. To predict if a patient will show up, a train and test set is required. The test set is created by randomly extracting one third of the total dataset. The remaining two third of the dataset is the training set. The target feature in the test set consists of “No-Shows” for 30.16% and 69.84% are “Show-ups”. In this research, imbalanced data should be taken into account. But this does not necessarily mean it will cause any problems.

In order to predict the show up at an appointment, five data mining algorithms are used. These algorithms are discussed explicitly in literature (e.g., Ye, 2003; Witten, Frank, & Hall, 2011). The first algorithm is the Naive Bayes. This algorithm uses the probability of the appearance within features to predict the show up of a patient (Witten, Frank, & Hall, 2011). The second and third algorithms are part of the decision tree. The algorithm “rpart” searches for the best feature to split the training set in two parts. This is repeated until no improvement can be made. The algorithm random forest picks a few random subsets out of the training set and builds decision trees on those subsets. While using this algorithm, there is chosen to set the number of trees to 50. After that, random forest combines these trees to create one decision tree (Ye, 2003). The fourth algorithm that is used to predict is the support vector machine (SVM). This algorithm selects a number of critical boundary instances from each class and builds a linear discriminant function that separates them as widely as possible (Witten, Frank, & Hall, 2011). The last algorithm that is used to predict is logistic regression. This algorithm gives a value between zero and one to every medical appointment in the test set. The higher this value, the more likely a patient will show up at his appointment (Witten, Frank, & Hall, 2011).

In Table 6 the percentages of those who are well predicted with the different algorithms are given. The first row of the table represents the features used to predict and the algorithms. If in the second row a box is colored, this means that the feature above that box is used to predict if a patient shows up at his appointment. For example, in the second row of Table 6, the results of the predictions using the different algorithms with the features age and waiting time is given. The last two columns of the table show the results of the logistic regression. The second last column shows the percentage of the well predicted appointments in the test set with a cut off at the 30.1616% quantile. The last column shows the percentage of the well predicted appointments in the test set with a cut off at the 50% quantile. This means that the lowest 30.1616% or 50% of the outcome of the logistic regression is put to zero (No show) and the rest of the outcomes are set to one (Show up). As can be seen, the highest percentage well predicted appointments is 69.84%. This percentage is the same as the percentage show ups in the test set. This is no coincidence. In most cases, no distinction is made within the features. The algorithms get the highest score predicting that everyone will show up. In that case, the percentage of well predicted appointments is 69.84.

Age	Gender	Day of the Week	Diabetes	Alcoholism	Hypertension	Handicap	Smokes	Scholarship	Tuberculosis	Sms reminder	Waiting time	Hour	Naive Bayes	Rpart	Random forest	SVM	Logistic regression (30.1616%)	Logistic regression (50%)
■											■		69.41	69.84	69.74	69.84	61.11	54.95
■	■	■											69.84	69.84	69.84	69.84	60.46	54.48
■	■	■								■	■		69.40	69.84	69.84	69.83	61.38	54.95
■	■	■								■	■	■	69.40	69.84	69.78	69.83	61.56	55.00
■	■	■	■	■	■	■	■	■	■				66.72	69.84	69.83	69.79	60.57	54.75
■	■	■	■	■	■	■	■	■	■	■	■		66.59	69.84	69.84	69.82	61.90	55.28
■	■	■	■	■	■	■	■	■	■	■	■	■	66.59	69.84	69.84	69.82	62.12	55.35
										■	■		69.33	69.84	69.84	69.84	61.00	53.84
										■	■	■	69.35	69.84	69.84	69.84	61.09	53.06
			■	■	■	■	■	■	■				66.38	69.84	69.84	69.81	41.51	41.51

Table 6*: The percentage of those who are well predicted. If the box is colored black, this means that this feature is used to predict if a patient shows up at his appointment. All training data is used here.*

3.1 Solving the imbalanced data

It appeared that the prediction showed unsatisfactory results. Mostly because predicting that everyone shows up, results in the highest percentage of correct predictions. A possible reason for this could be an imbalanced dataset. To solve this issue, there are a few suited methods. The first one is to double the no shows in the training set (Brownlee, 2015). This means that every appointment that has status “no show” now appears in the training set twice. After this is done, the target feature in the training set consists of “No-Shows” for 46.49% and 53.51% are “Show-ups”. When the no shows are doubled, they will appear more often and thus get more attention. In Table 7 the percentages of those who are well predicted with the different algorithms are given. As can be seen, the percentages in Table 7 are even lower than the ones in Table 6. So, for this training set, doubling the “no show” appointments is not a suitable method.

Age	Gender	Day of the Week	Diabetes	Alcoholism	Hypertension	Handicap	Smokes	Scholarship	Tuberculosis	Sms reminder	Waiting time	Hour	Naive Bayes	Rpart	Random forest	SVM	Logistic regression (30.1616%)	Logistic regression (50%)
■											■		62.79	62.30	61.33	61.13	61.12	54.93
■	■	■											59.45	55.59	59.17	58.96	60.45	54.47
■	■	■								■	■		62.12	62.30	61.41	61.95	61.36	54.93
■	■	■								■	■	■	62.25	62.30	61.62	61.89	61.56	54.98
■	■	■	■	■	■	■	■	■	■				65.99	55.59	61.32	55.51	60.58	54.76
■	■	■	■	■	■	■	■	■	■	■	■		64.47	62.30	62.29	62.71	61.93	55.28
■	■	■	■	■	■	■	■	■	■	■	■	■	64.43	62.30	62.00	63.48	62.14	55.33
										■	■		65.94	65.85	65.96	62.62	60.99	54.08
										■	■	■	65.74	65.85	62.13	62.49	61.05	53.34
			■	■	■	■	■	■	■				65.91	69.84	66.89	67.51	41.51	41.51

***Table** 7: The percentage of those who are well predicted. If the box is colored black, this means that this feature is used to predict if a patient shows up at his appointment. The no shows are doubled and added to the old training set to create a new training set.*

The second method to deal with imbalanced data is using the function SMOTE on the training set (Brownlee, 2015). The function SMOTE generates new appointments for the minority class, the appointments where patients do not show up. This is done by creating a new appointment and calculating the corresponding variables using the nearest neighbours of this class. This is called over-sampling. There is chosen to over-sample 500%. There is also a possibility to under-sample but this did not lead to better results. Therefore, the under-sampling factor is 100%. After this is done, the target feature in the training set consists of "No-Shows" for 54.55% and 45.45% are "Show-ups". This means that the training set even contains more No shows than show ups. After using the function SMOTE, the numbers are not integers anymore. A patient can now have 0.6 for the feature Diabetes. Even though this is unrealistic, this may help to predict if a patient will show up. In Table 8, the percentages of those who are well predicted with the different algorithms are given. As can be seen, the percentages in Table 8 are even lower than the ones in Table 6. Especially, the algorithms Naive Bayes and SVM have bad performance using this training set. So, for this training set, using the function SMOTE is not a suitable method.

Age	Gender	Day of the Week	Diabetes	Alcoholism	Hypertension	Handicap	Smokes	Scholarship	Tuberculosis	Sms reminder	Waiting time	Hour	Naive Bayes	Rpart	Random forest	SVM	Logistic regression (30.1616%)	Logistic regression (50%)
■											■		45.59	57.00	69.65	46.03	61.06	54.58
■	■	■											45.14	50.39	47.33	45.32	60.52	54.44
■	■	■								■	■		45.87	57.00	61.04	45.38	61.09	54.94
■	■	■								■	■	■	46.40	55.13	63.13	46.95	61.24	54.99
■	■	■	■	■	■	■	■	■	■				42.94	50.39	50.57	45.85	60.74	54.72
■	■	■	■	■	■	■	■	■	■	■	■		43.26	57.00	56.61	47.01	61.52	55.27
■	■	■	■	■	■	■	■	■	■	■	■	■	43.37	55.13	60.02	47.28	61.75	55.25
										■	■		30.47	54.72	44.69	47.68	60.84	54.06
										■	■	■	32.73	56.98	52.65	46.15	60.95	53.17
			■	■	■	■	■	■	■				40.71	40.87	41.20	40.40	41.61	41.61

Table 8*: The percentage of those who are well predicted. If the box is colored black, this means that this feature is used to predict if a patient shows up at his appointment. The training set is created by applying the SMOTE function over the original training set.*

The third method that can be used to deal with imbalanced data is to round the values calculated with the SMOTE function. As said before, the function SMOTE creates unrealistic values, for example 0.6 for the feature Diabetes. The results in Table 8 show that this did not lead to good results. To create more realistic values, there is chosen to round the values that came out of the SMOTE function. The target feature in the training set still consists of "No-Shows" for 54.55% and 45.45% are "Show-ups". In Table 9, the percentages of those who are well predicted with the different algorithms are given. As can be seen in Table 9, this method performs even worse than the method where the outcome of the SMOTE function is not rounded. For this training set, using the function SMOTE and then rounding the values is not a suitable method.

Age	Gender	Day of the Week	Diabetes	Alcoholism	Hypertension	Handicap	Smokes	Scholarship	Tuberculosis	Sms reminder	Waiting time	Hour	Naive Bayes	Rpart	Random forest	SVM	Logistic regression (30.1616%)	Logistic regression (50%)
■											■		45.59	48.87	49.50	45.14	61.06	54.96
■	■	■											45.26	48.87	45.54	43.55	60.52	54.49
■	■	■								■	■		45.91	48.87	50.94	44.89	61.09	54.94
■	■	■								■	■	■	46.37	48.87	51.97	46.10	61.23	55.01
■	■	■	■	■	■	■	■	■	■				44.99	48.87	48.58	43.91	60.74	54.72
■	■	■	■	■	■	■	■	■	■	■	■		45.52	48.87	51.18	44.67	61.53	55.26
■	■	■	■	■	■	■	■	■	■	■	■	■	45.69	48.87	51.81	45.24	61.76	55.25
										■	■		30.47	46.47	44.68	47.44	60.84	54.06
										■	■	■	32.39	50.77	49.31	46.56	60.95	53.16
			■	■	■	■	■	■	■				40.71	40.87	41.15	39.75	41.61	41.61

***Table 9**: The percentage of those who are well predicted. If the box is colored black, this means that this feature is used to predict if a patient shows up at his appointment. The training set is created by applying the SMOTE function over the original training set and then round these values.*

Despite the fact that these are all methods to deal with imbalanced data, they do not work well for this dataset. This may be because of the fact that the data does not differ enough for the patients who show up at their appointment and the ones that do not show up. During the data analysis, a test is conducted to check whether the age of the patients who show up at their appointment is significantly different from the age of the patients who do not show up at their appointment. This is also done for the waiting time. Both tests showed that these values significantly differ from each other. Nevertheless, if the age is plotted against the waiting time for both classes (Figure 8) no difference is visible. This can make it impossible to predict if a patient will show up at his appointment or not.

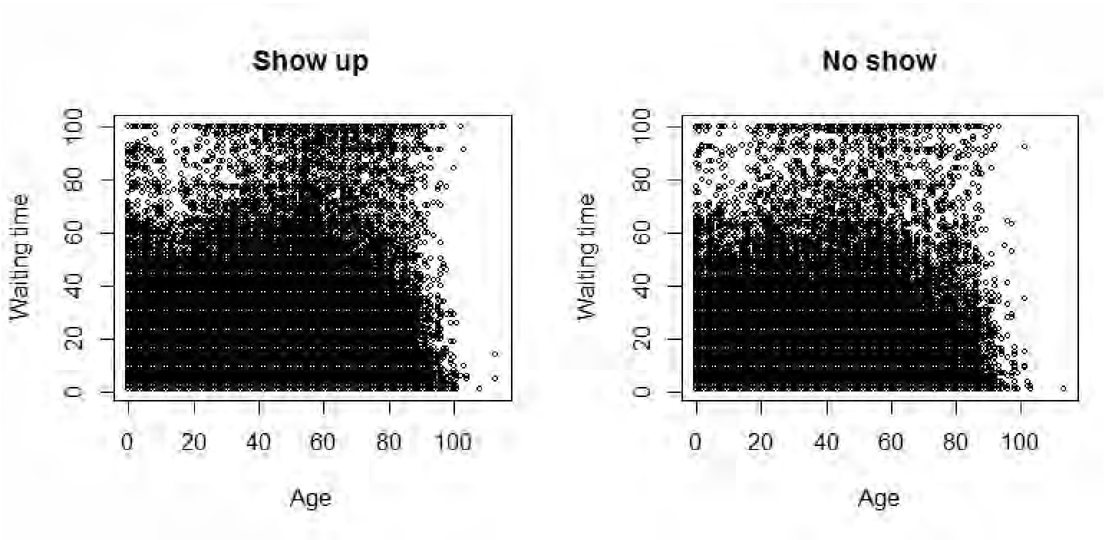

Figure 8: *The scatter plot on the left is shows the age and waiting time for the patients who show up at their appointment. The scatter plot on the right shows the age and waiting time for the patients who do not show up at their appointment.*

4. Recommendations

The aforementioned prediction model showed unsatisfactory results. However, other solutions exist to deal with gaps in the doctor's agenda resulting for a no show. The Bailey-Welch rule states that if for example the doctor starts at 8:00 am and one appointment takes 10 minutes, 2 appointments will be scheduled at 8:00 am. The next one is scheduled at 8:10 etc. In that case if someone does not show up, the doctor does not do anything and if everyone shows up then it expires for a maximum of 10 minutes (Guido & Koole, 2007). Because of the fact that in this case almost 30% of the patients do not show up at their appointment, more than one patient should be scheduled double on a daily basis.

Furthermore, the doctor or assistant could also ask for a reason for not showing up. By tracking all different reasons a potential pattern could be found. Also, instead of sending SMS reminders, gift cards could be rewarded for show ups (MGMA, 2009).

5. Conclusion

The purpose of this research was to predict if a patient will show up at his medical appointment or not and what to do if it is not predictable. During the data analysis, there is tested if the age of the patients who show up at their appointment is significantly different from the age of the patients who do not show up at their appointment. This is also done for the waiting time. Both tests showed that these values significantly differ from each other. Despite the fact that the data is imbalanced (70% show up and 30% no shows), predicting if a patient will show up should be easy. The algorithms used to predict the status of a patient are: Naive Bayes, Rpart, Random Forest, SVM, and logistic regression. None of these algorithms yields a result above 69.84%. This is exactly the percentage of show ups in the test set. Most of the models just put everyone on show up and then get a result of 69.84% correct. Because of the fact that the data is imbalanced, a few methods are tried to deal with this issue. These are the following:

- Double the medical appointments of the ones who do not show up at their appointment, this way more attention is paid to the minority class;
- Using the function SMOTE to create new data points, this way new data points are created to increase the number of appointments where patients do not show up at their appointment;
- Round the results of the function SMOTE, this way the new created data points still are realistic numbers that could have appeared in the original dataset.

Unfortunately, all of these methods did not lead to better results than 69.84% correct. These methods are all methods to deal with imbalanced data, but apparently success is not guaranteed.

Bibliography

Brownlee, J. (2015, August 19). *8 Tactics to Combat Imbalanced Classes in Your Machine Learning Dataset*. Retrieved July 12, 2017, from machinelearningmastery: http://machinelearningmastery.com/tactics-to-combat-imbalanced-classes-in-your-machine-learning-dataset/

Guido, C., & Koole, G. (2007). Optimal outpatient appointment scheduling. *Health care management science* .

Hoppen, J. (2017, Februari). *Medical Appointment No Shows*. Retrieved July 12, 2017, from Kaggle: https://www.kaggle.com/joniarroba/noshowappointments

Huisarten Praktijk Presikhaaf. (2017). *Niet verschijnen op afspraak*. Retrieved July 12, 2017, from Huisarten Praktijk Presikhaaf: https://huisartsenpraktijkpresikhaaf.praktijkinfo.nl/pagina/47/niet-verschijnen-op-afspraak/

MGMA. (2009, juli 9). *30 ways to reduce patient no-shows*. Retrieved juli 12, 2017, from MGMA: http://www.mgma.com/blog/30-ways-to-reduce-patient-no-shows

Witten, I. H., Frank, E., & Hall, M. A. (2011). *Data mining practical machine learning tools and techniques.* United States: Morgan Kaufmann.

Ye, N. (2003). *The handbook of data mining (Vol 24.).* Mahwah: NJ/London: Lawrence Erlbaum Associates.