Medical Appointment Show or No-Show Prediction

ISE/DSA 5103 - Intelligent Data Analytics

Team name: One Piece

Members: Anusha Saranam (113386735)

Sai Saketh Boyanapalli (113321570)

EXECUTIVE SUMMARY

Problem Statement

Patient's appointment classification, in brief is classifying whether a patient shows up at the time of appointment or not. This is very interesting problem as it is a common practise to make an appointment and miss it later for various reasons whatever it might be. The various factors involved might affect differently in the case of missing an appointment. The dataset hosted by kaggle consists of 110,527 medical records of public health care of the Espirito Santo State, Vitoria located in Brazil.

Major Concerns

This project is based on a Kaggle competition to predict if the patient shows up or not based on several characteristics such as gender, age, neighbourhood, diseases (like diabetes, hypertension, alcoholism) etc., The dataset consists of 100k rows of patient's appointments with 14 features per patient. We performed an exploratory data analysis on the dataset to get an initial gist of the data and features before continuing to model the problem. Despite no missing values in the data and very few outliers in 'Age' as it is the only continuous variable which include values like -2, -1 and 108, 113 etc. To deal with these, we've considered ages above 100 as outliers. We've done some feature engineering which seemed to increase the classification capacity of the model. Some problems in feature engineering are that we've decided to create a new variable appointment time, but all the data has date given with time as 00:00:00. Other problem is that all the data is factorial. The main issue is that proper weightage must be given to effectively evaluate the data. There's not much correlation between the variables too. There's very weak correlation between a few factors like Hypertension, Diabetes and Age pairwise which are not correlated with No show rate again.

Summary of Findings

Multiple data mining models are implemented on this dataset. They are Logistic regression, Decision tree, k-nearest neighbours, neural networks, Naïve Bayes and XG Boosting. Out of all these models, XG Boost gave the best results by far. The performance metrics accuracy, recall and specificity values for all the models are not so good but referring to the kernels in Kaggle, our results are better compared to those and as mentioned, XG Boost has the best overall results. Accuracy is the last preferred metric as the dataset is unbalanced with more 'No' and less number of 'Yes" for the No show rate which is the output predictor for the data.

Results

The best results were given by Extreme Gradient Boosting with the tuning parameters gamma = 0.001, c = 0.05 and max depth = 4. The results for the model were found out to be

recall = 0.989

specificity = 0.053

1. PROJECT UNDERSTANDING

Classification of Patient's appointment i.e. whether a patient shows up once an appointment is made plays an important role as there are other patients in queue. If a patient doesn't show up and that patient doesn't confirm whether they will be able to show up or not, wastes time and resources of themselves, healthcare management and other patients. In other words, the entropy of system is not minimum all the time. As healthcare is a primary resource to everyone, our aim in this project is to classify the no show rate (i.e. whether a patient shows up or not once an appointment is made) based on the factors available. To perform our analysis, we've chosen the dataset from kaggle. For testing the quality of our model, we've chosen parameters like Recall, Specificity and Accuracy.

2. DATA UNDERSTANDING

2.1 Data Description

The dataset provides data of 110,527 rows of medical appointments of the public healthcare of the capital city of Espirito Santo State, Vitoria located in Brazil and its 15 variables (characteristics) of each. The most important one if the patient show-up or no-show the appointment. Variable names are self-explanatory. Handcap is the total amount of handicaps a person has, it is not binary (0, 1, 2, 3, 4). Other variables include Age, Alcoholism, Diabetes, Hypertension, Scholarship some of which are binary. Problem Inconvenience is caused to other patients and doctors when an appointment booked doesn't show up. Out of 110,527 rows, there are 61744 unique patient ID's and all the appointment ID's are unique.

There are 14 features in the data. And created two more features.

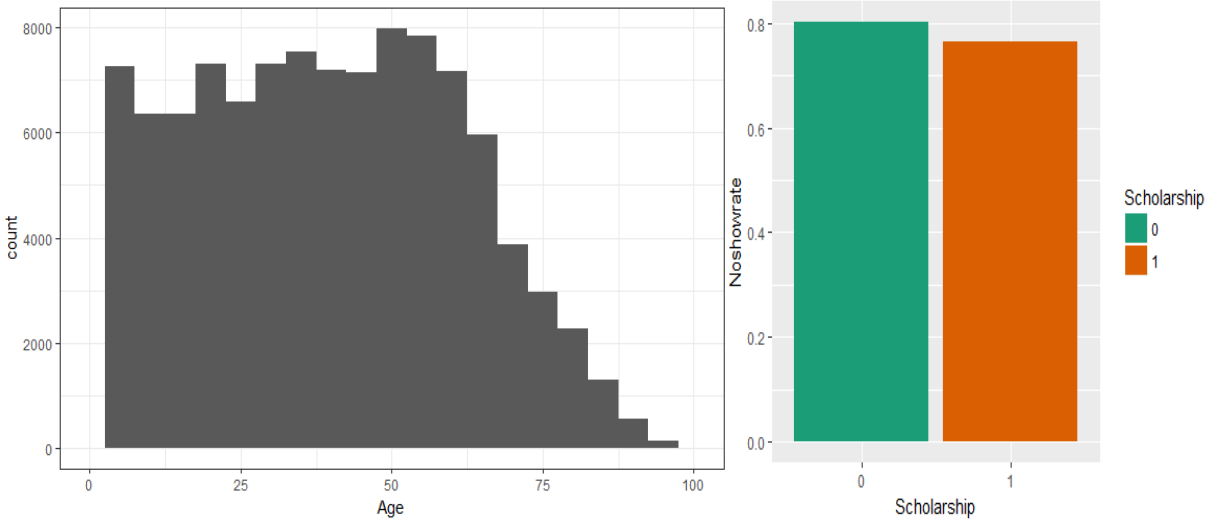
FEATURE	TYPE	FACTOR LEVELS
'PatientId'	Discrete	-
'AppointmentId'	Discrete	-
'Gender'	Categorical	2
'ScheduledDay'	Date-time	-
'AppointmentDay'	Date-time	-
'Age'	Continuous	-
'Neighbourhood'	String	-
'Scholarship'	Categorical	2
'Hipertension'	Categorical	2
'Diabetes'	Categorical	2
'Alcoholism'	Categorical	2
'Handcap'	Categorical	4
'SMS_received'	Categorical	2
'days_to_appointment'	Continuous	-
'weekday'	Categorical	7
'No.Show'	Categorical	2

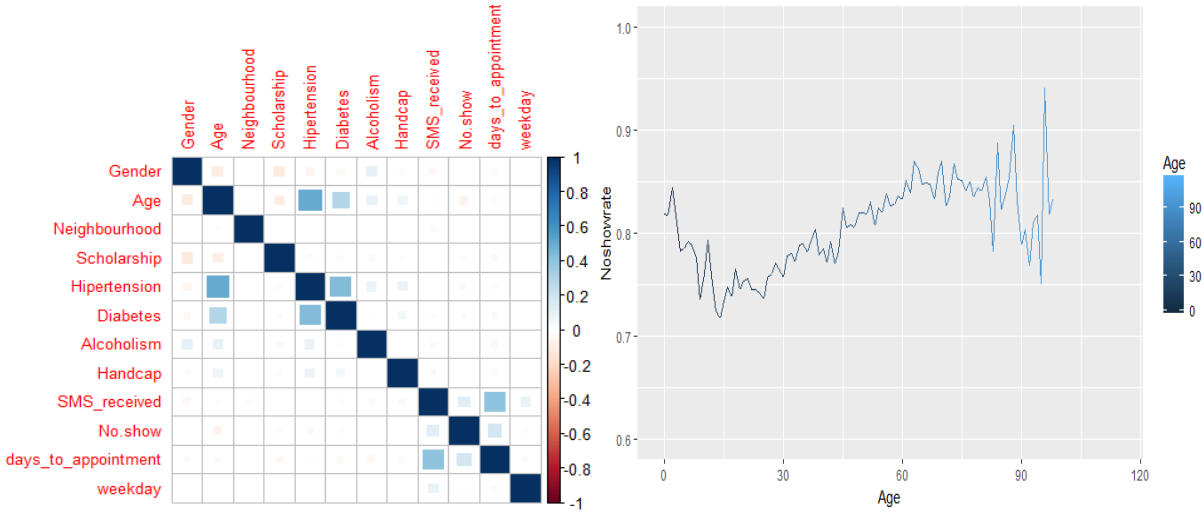
2.2 Data Visualization

This is very important first step in the model. There are 110K observations and 14 features in the dataset. Exploring age variable in the data we found that the age variable is slightly skewed, and it has got some negative values which is not acceptable for an age of the patient and there were few values above 100+ so, by removing them our data the variable is close to normal distribution. Created a new variable call no show rate which groups all the similar ages together in the data to find the ratio of noshow by sum of Noshow and show.

After the initial exploration of all the variables one thing was clear there was no clear relationship between the target and other variables in the dataset and the most of the variables we have are categorical, and these variables are highly skewed. But that is the case with the real-world data.

Here we can see that the no show rate is low for teenagers and then gradually increases as the age of the patient increases. Similarly, the no show rate is compared for all other variables available in the dataset and an interesting observation can be seen when it is compared to the patients with scholarships, there is a slight dip in no show rate about 5% for patients with scholarship.





3. DATA PREPARATION

3.1 Data Pre-processing

a. *Missing Values and Outliers:* There are no missing values in the data. The dataset is complete, and no imputations are carried out. Also, there are a few outliers in the data for the predictor 'Age'. Age includes values ranging from -2 to 113. For this purpose, we've treated negative values and values above 100 as outliers and removed them.

b. *Feature Engineering:* As per feature engineering, some features first created a feature called no show rate which is the obtained by grouping the variables and calculating the ratio of no shows to the total no. of values. This is done for all the values and compared the no show rate against all variable we have seen a relationship between no show rate and scholarship and age. People with scholarships tend to miss less than people without scholarships, as per age as the age increases the no show rate increases slightly and then decreases. There was a relationship between day of the week and the no show rate. Another feature was created using Scheduled Day and Appointment Day by taking the difference between two a new variable called days_to_appointment is created using lubridate package. Using the Appointment Day variable, a new variable called weekday is created which is a categorical variable that shows what day of the week the given appointment day falls. Exploring this variable, we see that no of appoints go up during Monday, Tuesday and Wednesday than steadily decrease during the weekends similarly people tend to show up more during the weekdays. Feature engineering will be key as making new features that are related to target will help the model in predicting. There were five levels for the variable handicap and the levels 1, 2, 3, 4 indicate that the persons have respective no. of handicaps but looking at the summary the levels other than 1 do not have many observations and it is highly skewed so, we have converted levels 2, 3 and 4 to 1.

3.2 Data Splitting

From the source of this dataset, there is no specific test to check the model. We've tried different split ratios like 80:20, 70:30 and 50:50. But because of the unbalanced dataset, we've decided to use 50:50 for train and test data. We also used random sampling for effective modelling.

4. MODELLING

This data set is huge so, it will be quite challenging to tune the models so that the algorithm predicts with high recall, specificity and accuracy and runs efficiently. Tuning the parameters effectively will be very important. some of the features by themselves are not related to the target, the goal will be to come up with the features that are related to the target and improve the performance of the algorithm. Preparing the data to be suitable for the classification algorithms, we have performed feature selection manually by removing features PatientId, AppointmentId, ScheduledDay, AppointmentDay. And

incorporated above mentioned features into the model. The data is validated using a split sample method which randomly divides the data into two sets training and test set of which the training set is used to train the algorithm and the test set is used to evaluate the algorithm. To evaluate the performance of the algorithm accuracy, recall & specificity were used as evaluation metrics mainly recall and specificity are important as the problem is of classification type.

Created models using logistic regression, Decision tree, K nearest Neighbour, Neural net, Naïve Bayes and Extreme Gradient Boosting (XG Boosting) to compare the results.

4.1 Brief Explanation of Modelling Algorithms:

4.1.1. Logistic Regression Model

For finding out the relationship between the response variable No show rate and other independent variables and the normal distribution is taken care of. However, since the data is not linear, this model performed poorly yielding poor results. With this, we can conclude that the data is not linear and varies non-linearly with the data.

4.1.2. Tree Methods

Decision Trees: Unlike linear methods, trees map non-linear relations quite well. This dataset uses continuous variable decision trees (classification problems). Decision trees also help in variable importance. Decision trees have low bias and high variance, so this might cause over-fitting from the model, we calculated the complexity parameter, cp which turned out to be 0.01 from the minimum x error (x error is the cross-validation error).

4.1.3. KNN

In this K-closest observation are defined as the ones with the smallest Euclidean distance to the data point under consideration. For prediction, in classification, it takes the weighted average of the values of the K-neighbours based on the predictors. This model takes a lot of time to run because of the large data. The model is tuned for different k values, but the best value is found out for k value around 318.

4.1.4. Neural Networks

A Neural Network is a classification technique designed based on the biological working of the nervous system. For this dataset, we have used the nnet method to train the neural networks. We must predict the binary output for no show rate. Data used is scaled & centered. Tuning parameters include size and decay for the model.

4.1.5. Naïve Bayes

Naïve Bayes is a better algorithm for binary classification works better for real world problems. It is based on the baye's theorem with an assumption for independence among predictors. To tune the algorithm, there is only one parameter 'size'. This is the second best model after XG boosting model which is by far the best model as it works better for very large datasets.

4.1.6. Extreme Gradient Boosting (XG Boost)

It is an accurate and effective procedure used for classification. It combines the outputs from weak learners and creates a strong learner, thus improving the prediction power of the model. It is used to impart additional boost to accuracy. This is the final model used for our problem. This model has the best recall and specificity.

4.2 Validation

Validation is an approach for evaluating the performance of the algorithm when we input data that's different from the data that we have used in our algorithm for training. In this technique we split the data into two parts testing and training sets. There are many approaches to splitting the data into two sets. After the data is split the training set is used to train the algorithm and the test set is used to predict the model. If we do not perform validation and evaluate the model with same sets of data one might end up with a model that is overfitting the data.

Here we are using sample split to split the data into two sets. Size of the training set is 50% of the data and the test set is remaining 50% of the data. Sample split uses an approach where it comes up with an array of Booleans for the indexes randomly depending on the size of test data, these values can then be run against the data to create training and test sets.

4.3 Evaluation

In machine learning the performance of the algorithm is measured using different metrics which is very important in determining the right model for the problem in hand and evaluating the performance across various algorithms for different sets of data is important to get an overall picture for the performance of the algorithm.

The problem we are dealing with is a classification problem. For every algorithm we generally measure the accuracy of the algorithm and evaluate the performance using this metric.

$$Accuracy = \frac{Number\ of\ labels\ predicted\ correctly}{Total\ number\ of\ predictions}$$

Which is a good evaluation metric for a regression problem. But is not robust against a classification problem, the problem here being that if the algorithm just predicts every customer to miss the appointment we might end up with an overall accuracy of 70% which is considered very good. T=but the problem is we want to correctly classify them.

Two robust evaluation metrics to be used with a classification problem is Precision, Recall and Specificity.

$$Precision = \frac{True\ Positive}{True\ Positive + False\ Positive}$$

In case of precision, it is how many times the algorithm was able to correctly predict that one is going to miss an appointment of all the times it said the person is going to miss. In simple terms, it is how accurate the algorithm is when it says the person is going to miss.

$$Recall = \frac{True\ Positive}{True\ Positive + False\ Negative}$$

The recall is the ability to correctly recognize if a person misses an appointment. Which is very important in our case as we want to correctly predict if a person is going to show up or not show up.

The Specificity of the model which is True Negative Rate is the ratio of correctly identified negatives and total negatives. The formula of Specificity is

$$Specificity = \frac{True\ Negative}{True\ Negative + False\ Positive}$$

5. RESULTS

5.1 Results

The results of all the models for this project are tabulated below

MODEL	TUNING PARAMETERS	ACCURACY	RECALL	SPECIFICITY
Logistic Regression	-	0.79	1	0
Decision Trees	cp =0.01	0.697	0.942	0.044
K Nearest Neighbours	k = 318	0.73	0.96	0.06
Neural Networks	decay = 0.1	0.70	0.93	0.12
Naïve Bayes	smooth = 0.001	0.714	0.97	0.052
XG Boosting	c = 0.05 max_depth = 4 gamma = 0.001	0.8	0.989	0.053

5.2 Conclusion

The aim of the project is to predict if the patient is going to show up or not at the time of the appointment, went through the process of data analysis by performing exploratory data analysis, creating new features, dealing with outliers and variable transformations, data validation and implementing them into the model. The results seem to be promising with the kind of data we have for the problem given the predictors are skewed and most of the predictors are categorical. Extreme Gradient boosting was the best performer when compared amongst other models. It was challenging to come up with correct evaluation metrics for the problem it is easy to look at false metrics that give promising results.