**DSC 789-001-spring 2022**

**Strategic Capstone Projects**

**Xuan Liu**

**Weijie Zhao**

**5/16/2022**

**Final Project Report**

In the medical industry, hospitals from around the area are negatively affected by no-show appointments. No-show appointments have been a serious problem that occupies medical resources and wastes a large amount of money and time; furthermore, missed medical appointments even impact the efficiency of medication to save patients’ lives (Qureshi et al., 2021). In fact, the number of patients who miss appointments is gradually increasing every year. As for hospitals, the system of medication has to encounter heavy challenges and potential risks. As for patients, healthcare services will not be convenient as usual.

Therefore, it is important to avoid the risks from no-show appointments, and utilizing an optimal medical appointment system is able to reduce the increased idle time, overtime, and waiting time that other patients and hospitals face in order to increase revenue and improve productivity and quality of care (Batool et al., 2020).

What we have to do in this project is create machine learning models to predict clients’ appointment behaviors and help this healthcare center to timely adjust their own appointment system. This would greatly improve health care delivery and efficiency for the center and patients. In addition, after this project, we should answer these research questions down below.

1. Can we predict the variable of interest (response variable) using data analytical

techniques?

2. What are the important features in order to predict the response variable?

3. What is the importance of the features?

4. Can we predict the probability of each class (classification problem) for each sample using data analytical techniques?

5. How do the different sampling techniques affect the prediction of the response variable?

6. How do the different feature selection methods affect the prediction of the response variable?

7. How does the prediction performance change for more complex analytical models?

8. What strategy outline do you propose to increase the performance of the specific problem (healthcare, banking, service) you are working on? Or how do you apply the data analysis that you have into practice in order to make data-driven decisions for the business you are studying?

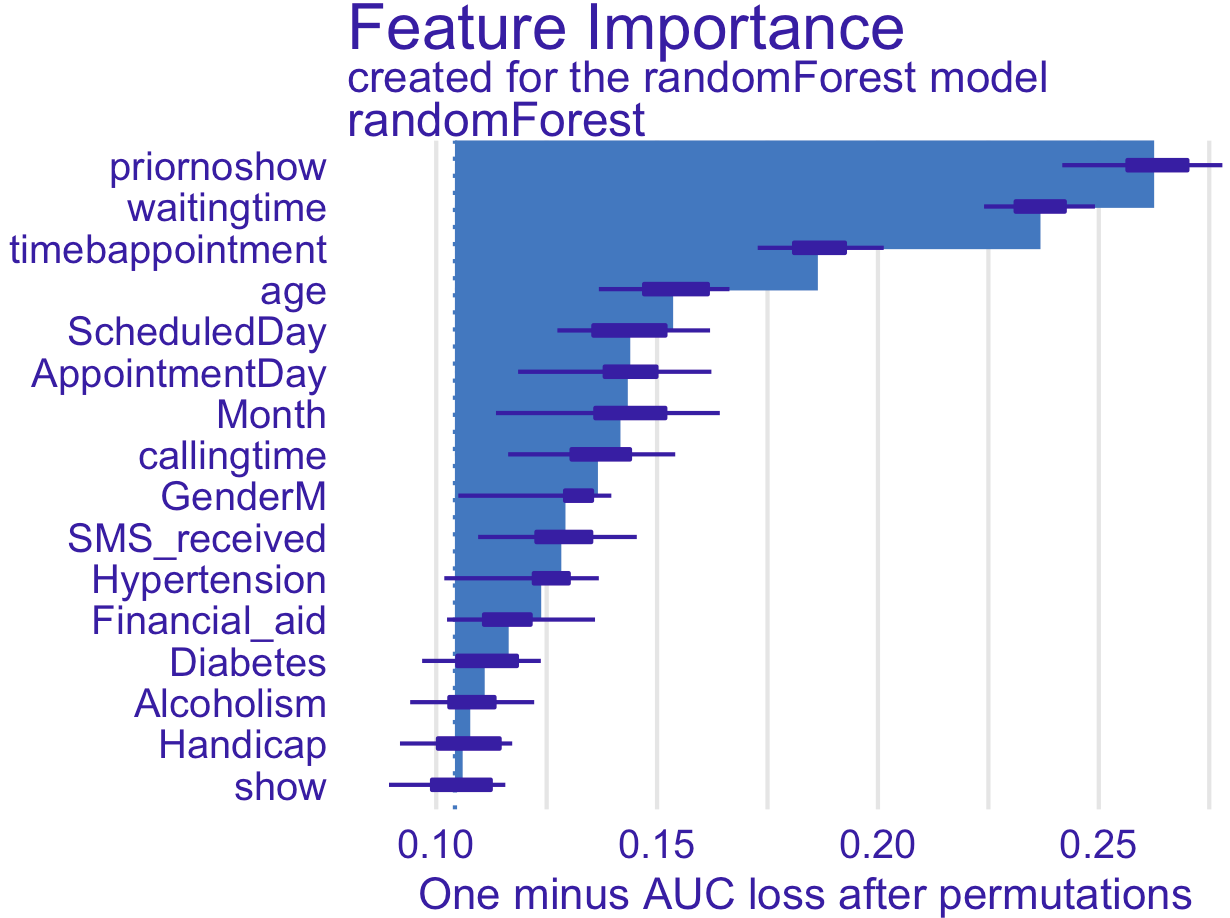
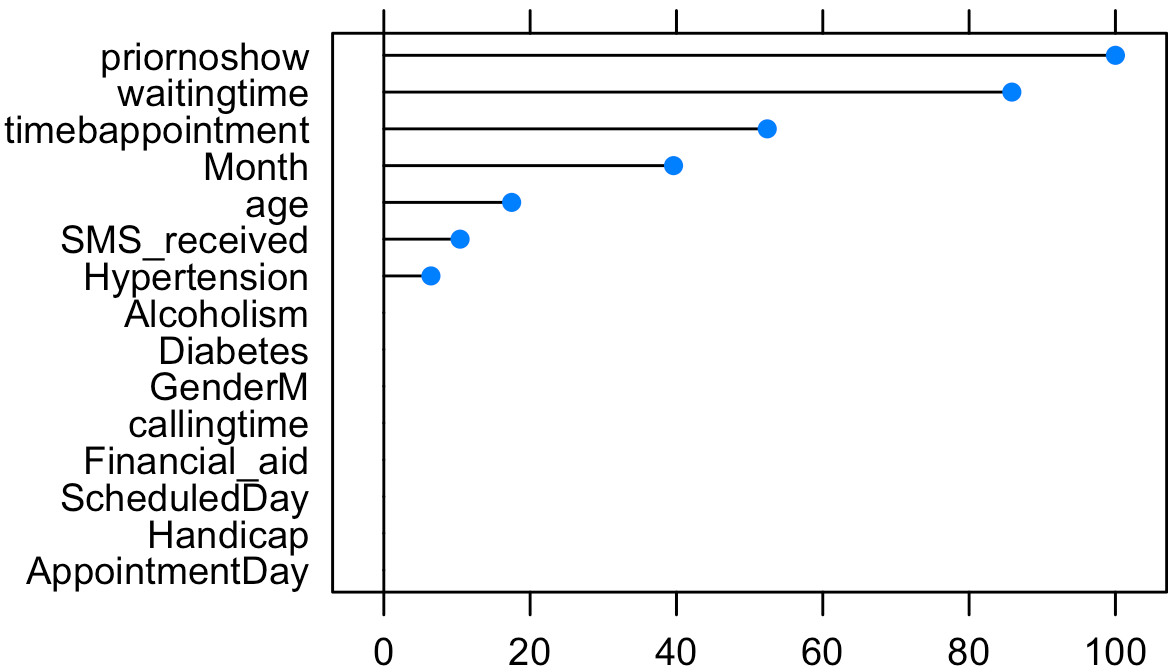
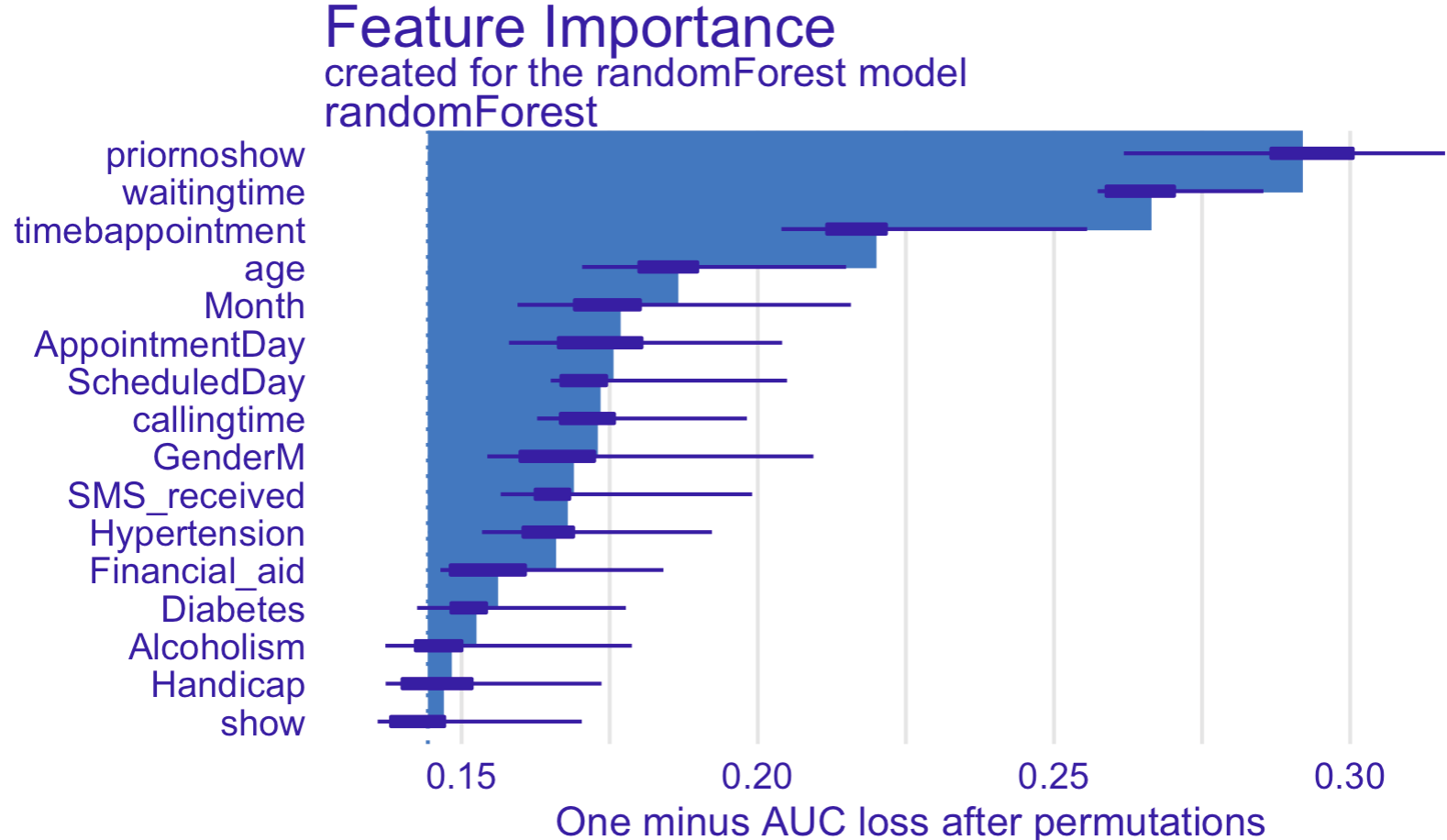
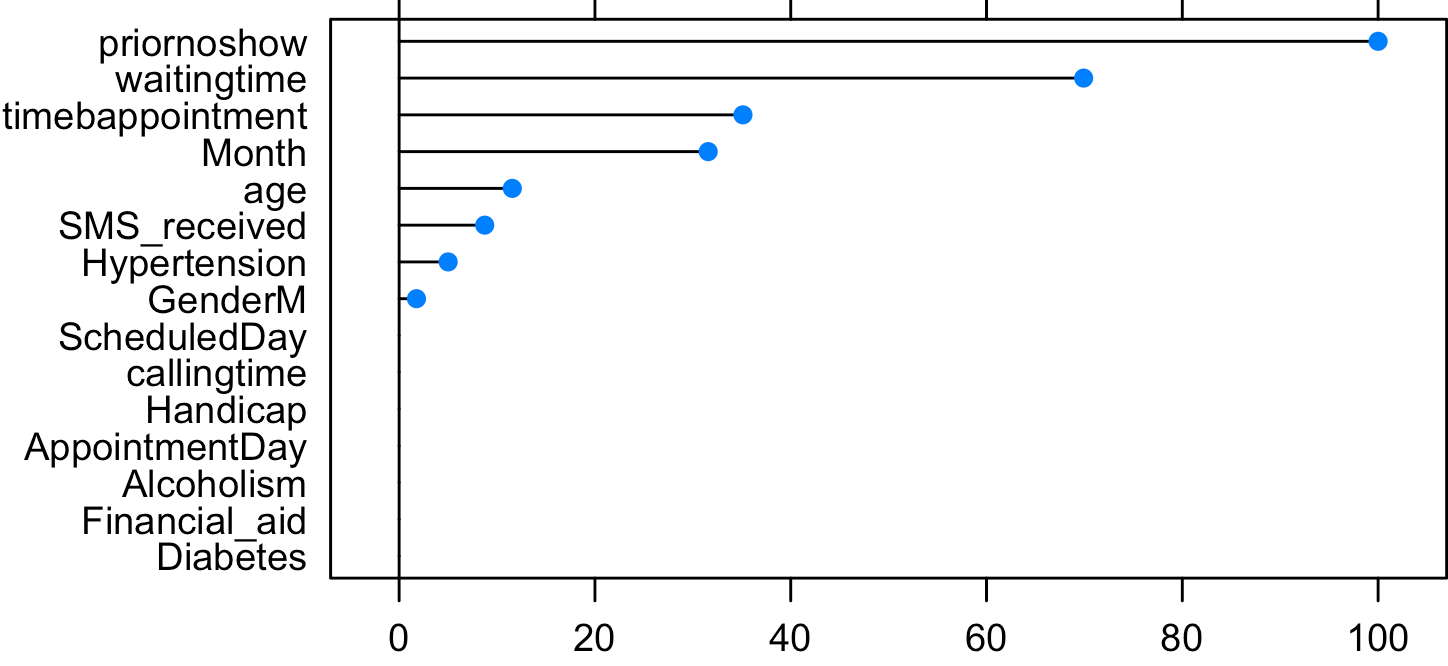
Moreover, for the methodology section in this project, firstly, by using R, we knew more about the descriptive statistics of this dataset. Then we created graphs for each variable and input variables with output variables in our dataset through Tableau to know the distribution of them and then better analyze those data later on.

Secondly, it is the data preparation phase. We need to make our data more organized by applying multiple methods of R. At the beginning, we removed redundant features and checked missing values. Dealing with the near zero variance variables was the next one. We changed categorical variables into numerical variables using integer and one-hot encoding. Then binning variables Age, Calling\_time, Waiting\_time, Time\_b\_appointment and Prior\_noshow by applying equal frequency or width was another strategy for the noisy data. Furthermore, it is vital for us to detect the outliers in the dataset. Clustering was the method we chose to do outlier detection and picked the top 1% outliers to delete.

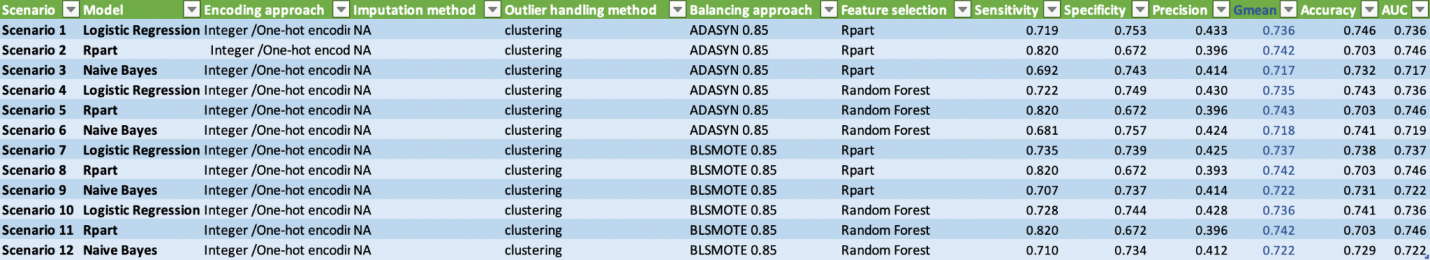
Thirdly, it is the modeling sector. we splitted the data into train and test sets. The percentage is 75% for training and 25% for testing. Additionally, we applied Rpart and random forest models for the feature selection part. Then utilizing the “imbalance” package to balance the data was the next part. We used two different resampling methods(ADASYN and BLSMOTE) to compare multiple scenarios later. We applied three different classification models for our dataset which are Logistic Regression, Rpart, and Naive Bayes. Therefore, with two different resampling and feature selection approaches as we mentioned before, there are totally 12 machine learning models we have. Besides, in order to evaluate the models we got, we utilized the confusion matrix and a few classification evaluation metrics(sensitivity, specificity, precision, Gmean, accuracy, AUC).



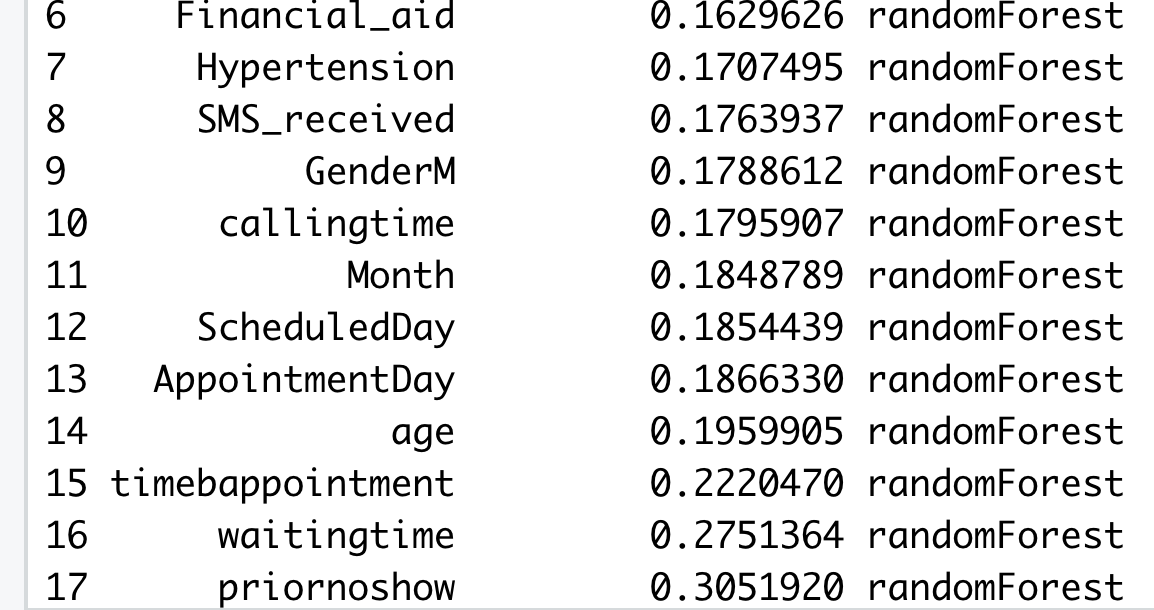
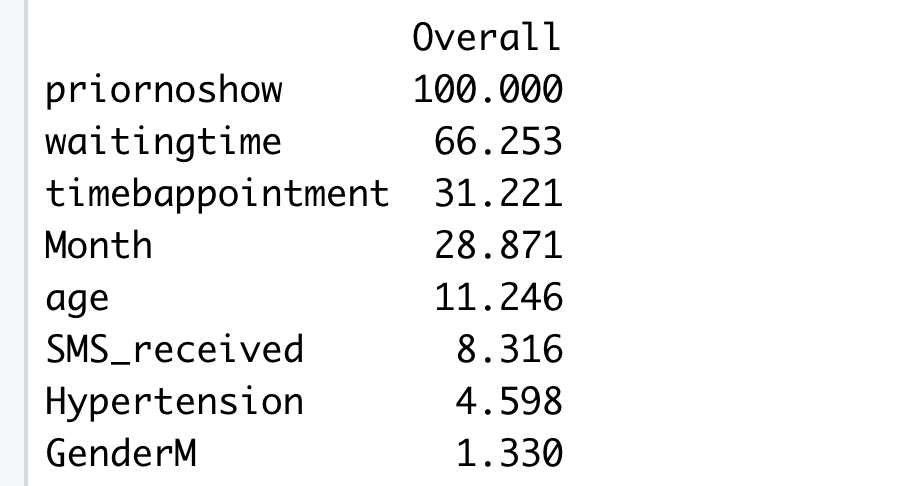
What’s more, for the results of this project, we had a couple of conclusions. According to these four figures of feature importance we showed below, the most important features for the no show appointment dataset are prior no show, waiting time, time between appointment, age, Month, SMS received.



In addition, in the results table we created down below, we could see different evaluation metrics for each model. For sensitivity, it is the true positive rate(TP/(TP+FN)) and specificity is the true negative rate(TN/(TN+FP)). Precision is TP / (TP+FP) and G mean is the geometric mean of sensitivity and specificity. Then accuracy is the result of (TP+TN) / (TP+FN+FP+TN). AUC is the area under the ROC curve. Moreover, compared the feature selection sector of 12 models we have now, we can tell that Rpart has a better performance for this dataset since their matric scores like sensitivity, specificity, precision, Gmean or AUC are better than random forest in general. Besides, for the balancing approach, we found that BLSMOTE is better than ADASYN. And for the model part, logistic regression is the best model we chose because their matric scores are higher and more stable than the other two models in general.



Furthermore, considering the research questions we showed before, we can predict the target variable using data analytical techniques like machine learning algorithms we did. We knew that Prior no show, waiting time, time between appointment, age, Month, SMS received are the important features in order to predict the target variable from the analysis process. There are two different feature selection methods and we had 4 feature selection results. For their importance, you could know from two examples down below.



We could also predict the probability of each class (classification problem) for each sample using a data analytical method like the confusion matrix we got. Besides, different sampling techniques and feature selection methods can affect or improve the prediction of the response variable, which we can tell from the result table above. A proper sampling technique or feature selection method may give us better models. Then, in 12 scenarios, we used the Rpart model which is a more complex model. However its performance is not that good. Hence, more complex analytical models don’t mean that the performance would be better. In the end, we got the most important features. It means the healthcare center could focus on those factors and have them into practice in order to make data-driven decisions.

All in all, it is the discussion and conclusions section. In general, the 7th model(logistic regression model with BLSMOTE resampling method and Rpart feature selection) is the best model we have so far. Additionally, prior no show, waiting time, time between appointment, age, Month, SMS received are the most important factors for this no show appointment problem. Therefore, this center should focus on any factor they could have control like waiting time and SMS received. They should try to reduce the time of patients waiting for the appointment and improve the efficiency of their system. And having reminder messages or calls before appointments is also essential for the no show appointment issue. We believe that this issue would be better corrected if they could do some changes on these 2 factors.

**References**

Qureshi, Z., Maqbool, A., Mirza, A., Iqbal, M. Z., Afzal, F., Kanubala, D. D., Rana, T., Umair, M. Y., Wakeel, A., & Shah, S. K. (2021). Efficient prediction of missed clinical appointment using machine learning. *Computational and Mathematical Methods in Medicine*, *2021*, 1–10. https://doi.org/10.1155/2021/2376391

Batool, T., Abuelnoor, M., Boutari, O. E., Aloul, F., & Sagahyroon, A. (2020). *Hub: Predicting hospital no-shows using machine learning. 2020 IEEE International Conference on Internet of Things and Intelligence System (iotais): 10.1109/iotais50849.2021.9359692*. Sci. Retrieved April 2, 2022, from https://sci-hub.hkvisa.net/10.1109/iotais50849.2021.9359692