ALY 6040: Data Mining

Final Project Report

Darshil Oza, Nilay Anand, Rohit Meena, Vamshi Paili, Yash Tadiyal

College of Professional Studies, Northeastern University

Prof. Justin Grosz

May 8, 2022

Introduction

The goal of this study is to predict if a diabetes patient will be readmitted to a hospital and the factors affecting readmission. This will be done in perspective of an insurance company trying to gauge the cost of treatment and reduce the chances of having to pay for the treatment or be better prepared for the eventuality.

Data Cleaning

The data set contains 100,000 rows and 50 columns containing data about diabetic patients who underwent treatment for diabetes. The data set on first inspection showed no missing values, on closer inspection there were some columns that contained "?" symbols instead of data. The first step we took was dropping columns with too many unknown values or columns that were irrelevant to the problem at hand. The first columns to be dropped were medical_specialty, payer_code, patient_nbr, encounter_id and weight. Next, rows with the question mark symbol were dropped after looking at the percentage of missing values and if there were important. After dropping, the data set was left with 98,000 rows. So, dropping the data did not affect the quality of the data as only 2% of the data was dropped. Next, admission_type_id, discharge_disposition_id and admission_source_id columns were dropped since they are irrelevant and contain identification data which is not useful for the predictive model. The column containing gender values was converted to female, this allowed us to represent gender in form of Boolean values making it easier to ingest during model training. The race column was converted to dummy variables which created individual columns for each race in form on Boolean values. The data set also contained 23 columns for different medication that were given to the patient. Some of these were used in most cases and rest were rarely used. These drugs were divided into regular and scarce types, and they were dropped since we are focusing on the diabetes medication as the primary source of the treatment in determining the chance of readmittance. The column we looked at next was the one we were trying to predict, the readmitted column contained three types of values: No, <30 and >30. The <30 and >30 represents patients that were readmitted less than 30 days and more than 30 days of treatment. These were encoded as 1 and the No values were encoded as 0. This allowed the prediction variable to be encoded in binary form. Two more columns, change and diabetesMed had data in form of “Ch”/”No” and “Yes”/”No” respectively. These columns for changed to Boolean values. The age column contained age in ranges of ten years, and these were replaced by the upper end of the range for each row which gave us age by the decade. Finally, 5 columns: diag_1, diag_2, diag_3, max_glu_serum, A1Cresult were also removed as they had unexplained values and were irrelevant to the building the final model. We visualized the correlation between the variables that were left and got the figure shown in Figure 1. As seen in the figure the number of medications and number of lab procedures have a positive correlation with time spent in the hospital which was expected. There was no other significant relation between the variables. Next, we look at few visualizations to find any useful insights in data before modelling.

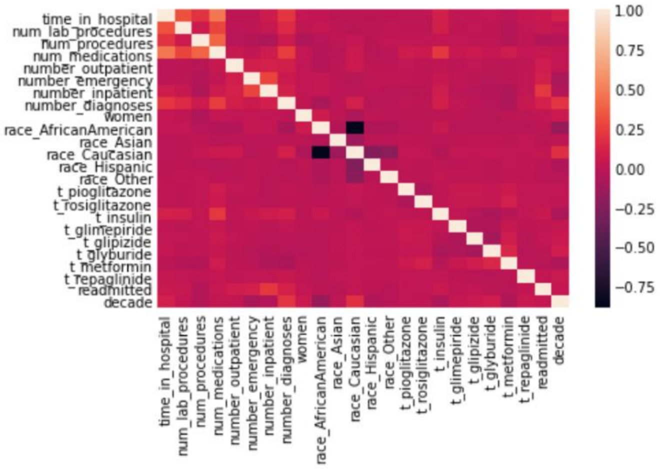


Figure 1: Correlation matrix

Exploratory Data Analysis

We looked at variables we'll be using for the final model. The plot can be seen in Appendix A Figure 1. We look at this data to get a deeper understanding of the distribution of variables and whether the once we selected would be fit for modelling. Going through each graph one by one, we see that in the decade graph contains most values in the range of 50 to 90, most of the patients dealing with diabetes lie in that range. The time spent in the hospital lies mostly in the range of 1 to 4 days. The number of lab procedures follow the pattern of a normal distribution and seems almost evenly distributed across. Most values were in the range of 40 to 70, also the maximum number of patients had only one lab procedure. Most patients had no procedures done and about 40,000 patients had either 1, 2 or 3 procedures done. The number of medications given per patient had the greatest number of patients taking 6 to 20 medicine which is good range of possible and well distributed, which indicates the variable might be a useful indicator of readmittance. Number of outpatients, inpatients and emergency are highly skewed towards zero and probably will not be useful and have a lower representation while modelling. The last graph for number of diagnoses is highly skewed towards 9 diagnoses but is also evenly distributed in the range of 5 to 8 diagnosis. This column might show up as useful or significant in the models.

We looked at out prediction variable next, the graph for count can be seen in Appendix A Figure 2. The readmittance looks evenly distributed across the data set with 53% no and 47% yes. The model we build using this should be able to predict both cases with equal accuracy albeit seeing the distribution of variable, the accuracy maybe low. Next, we looked at the effect of time being spent in the hospital on readmission. The plot can be seen in Appendix A Figure 3. It shows that readmittance chance increasing with more time spent in the hospital, it indicates that there might be some positive correlation between the two. The plot between Age of patient and readmission as seen in Appendix A Figure 4, indicates that most patients lie in the higher age ranges and there seems to equal distribution of being readmitted and not readmitted across all age groups. The effect of race can be seen in Appendix A Figure 5 and most cases are of the Caucasian race, although the African American seems more likely to be readmitted if they get treated for diabetes. Appendix A Figure 6 shows if the number of medications affects readmission and there seems to no effect as the distribution is even between the two. The plot between gender and readmission as seen in Appendix A Figure 7 shows an uneven distribution between the two genders. More number of females have cases of diabetes, and their chance is readmittance is also higher. They pose a higher risk for the insurance company. The plot shown in Appendix A Figure 8 shows the change is a patient that is given treatment vs readmittance, people that seem to respond to treatment were readmitted less. For the insurance company, people responding under treatment as less of a readmittance risk. As seen in Appendix A Figure 9, most patients are given diabetes medication and distribution between readmitted and non-readmitted seems equal. The final plot between Glucose test serum test and readmission in Appendix A Figure 10 shows no results of significance.

Modelling

The model we chose were all classifiers as we have a binary variable as our target. The algorithms we chose to compare are simple logistic regression, tree based and tree boosting ones.

1. Logistic Regression

Logistic regression analysis is valuable for predicting the likelihood of an event which in this case is the readmittance of a patient . It helps determine the probabilities between any two classes and is a decent first model to implement as the user can set their expectaions based on the results of this model. The dataset was split into two parts in the ratio of 3:1 (75% & 25%) for training and testing data, respectively, for all the models. First, we plotted the statistical model of the dataset to gain some initial insights based on the p-values and absolute coefficient values. As a result, we found the racial ethnicity columns in the dataset to be of the most importance compared to the rest of the columns. However, this result is not entirely true. The other important columns were diabetes medication and the number of patients being admitted. The accuracy of the logistic regression model is 61% and the recall rate is just 40%. The number of false negative cases is 6855. However, the number of false positive cases is 2686 , which is the best value among all the models.

2. **Decision Tree**

We decided to use the decision tree classifier in this project as it provides flexibility to use a variety of feature subsets and decision rules at various stages of the classification procedure. A big advantage of using a decision tree is that it pushes you to think about all the possible consequences of a decision and tracks you along each route that leads to a conclusion. It generates a complete analysis of the effects along each branch and flags decision nodes that require more investigation. The accuracy of the decision tree model was 62% and the recall rate was 50%. The number of false negative cases is much lower compared to linear regression at 5776. However, the number of false positive cases is relatively lower at 3629. We set the maximum depth of the tree to five and were able to figure out the most important columns of the model for evaluation by plotting the entire tree. The most important columns of the model, in decreasing order of their importance, are the number of diagnoses, decade, diabetes medication, and number of laboratory procedures.

3. **Random Forest**

We decided to implement the random forest model in the project as it can handle binary features, categorical features, and numerical features. As a result, there is very little pre-processing that needs to be done and the data does not need to be rescaled or transformed. Random Forest has methods for balancing errors in the class population in unbalanced data sets by minimizing the overall error rate. And since there is data in some of the columns in the dataset that is indeed unbalanced, as we identified during EDA, it will be able to handle it well. The accuracy of the random forest model is 60% and the recall rate is 51%. Moreover, we were able to determine the most important features or columns of the model through feature importance. The five most important features of the model, in decreasing order of their importance, are the number of laboratory procedures, number of medications, time spent in hospital, decade and the number of diagnoses. The number of false negative cases was 5566, which is better than both the previous models. The number of false positive cases was 4210, which is worse than both the logistic regression and decision tree models.

4. Gradient Boost

We used the gradient boost model as, in most cases, the predictive accuracy of its model is the highest compared to the rest of the models, irrespective of whether it is logical or practical to use the model. Furthermore, it can optimize on various loss functions and offers several hyper parameters tuning options, making the function fit very flexible. The accuracy of the gradient boost model is 62% and the recall rate is 47%. We did not use feature importance to find the most important columns used by the model to evaluate the results because the gradient boost model works on optimizing the resultant curve to satisfy the data points rather than following a logical or mathematical method of evaluation. The number of false negative cases was 6112, which is worse than both random forest and decision tree models. The number of false positive cases was 3222, which is better than the previous two models.

5. ADA Boost

We implemented the ADA Boost model in the project because it is adaptive in the sense that subsequent classifiers created prefer examples misclassified by earlier classifiers. Furthermore, the model encapsulates several non-linear relationships, resulting in improved prediction accuracy on the problem of interest. The accuracy of the ADA boost model is 62% and the recall rate is 45%. We did not use feature importance to find the most important columns used by the model to evaluate the results because the ADA boost model works on optimizing the resultant curve to satisfy the data points rather than following a logical or mathematical method of evaluation. The number of false negative cases was 6325, which is not better than random forest. The number of false positive cases was 3051, which is better than both the previous models.

ADA Boost and Gradient Boost both learn sequentially from a small pool of learners. The additive model of these weak learners yields a strong learner. The major goal here is to learn from the flaws at each stage of the iteration. ADA boost is primarily concerned with 'voting weights,' whereas gradient boosting is concerned with 'adding gradient optimization'. The accuracy between models is not significantly different. So, we chose minimum the false negatives as the metric for comparison. In context of an insurance company, it's goal would be to minimise false negative as the cases in which the model classifies a patient as negative for readmission and then they are readmission creates a liability for the company which they were not prepared for, and it becomes as unforeseen expense to be paid out.

The best model for our use case is the Random Forest model as it had a comparable accuracy to other models, but the false negatives were the lowest of the bunch. We also looked at the features that were important to the model and the number of lab procedure, number of medications, time spent in hospital and the age of the age are the four highest features of importance. The model results can be seen in Appendix B

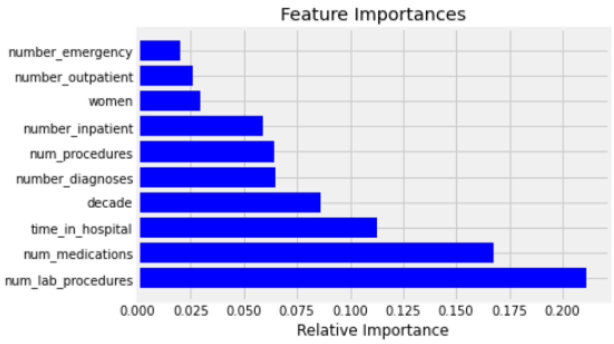


Figure 2: Random Forest feature importance

Conclusion

According to our preliminary findings, diabetes is more commonly found in older people, specifically those aged 50–90, and it affects women more than men. Furthermore, medicine is more beneficial at older ages than at younger ones. The rate of readmission is greater in older age groups than in younger age groups. This might be attributed to increased health difficulties as the body ages. Younger people recover more swiftly and their bodies repair more quickly. The success of diabetic treatment varies from case to case, but if patients are not given the right medication, readmission rates can skyrocket.

The best-case scenario for a health insurance company would be that their client does not have to get readmitted into the hospital because it would increase the medical bills of the client that the company would have to pay. We have concluded that early identification of diabetes through tests and then getting the proper medication significantly lowers the chances of getting readmitted to the hospital. Therefore, the insurance company should provide an incentive to their clients to regularly get tested for diabetes. This process has an initial cost for the company, but it will yield exponential revenue in the long run. The insurance company could provide a credit reward system to their clients, in which clients could earn credit points that they could use to claim a product or a membership or subscription. The company could further improve the chances of their clients not getting readmitted by offering things that further help improve their health. The list of products the company could offer could include salads, vegetables, fruits, medicines, etc. The company could also offer the clients a membership to gyms, spas, bike rentals, etc., that could help their physical health.

Appendix A: EDA Visualizations

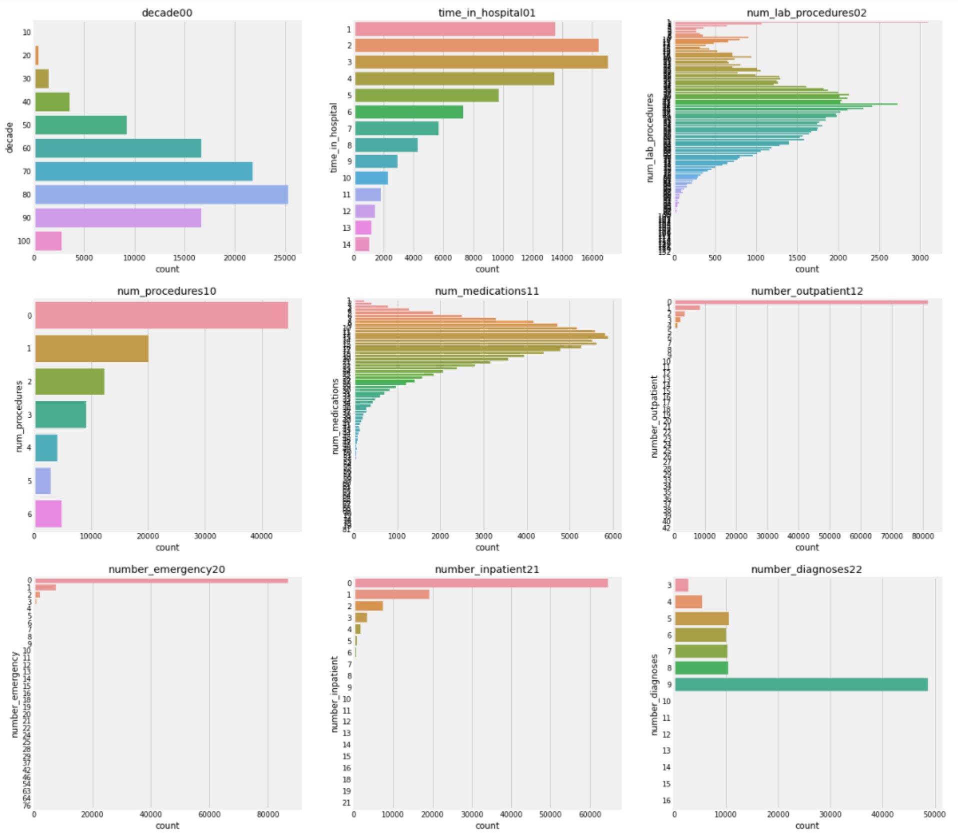


Figure 1: Visualization of distribution of variables

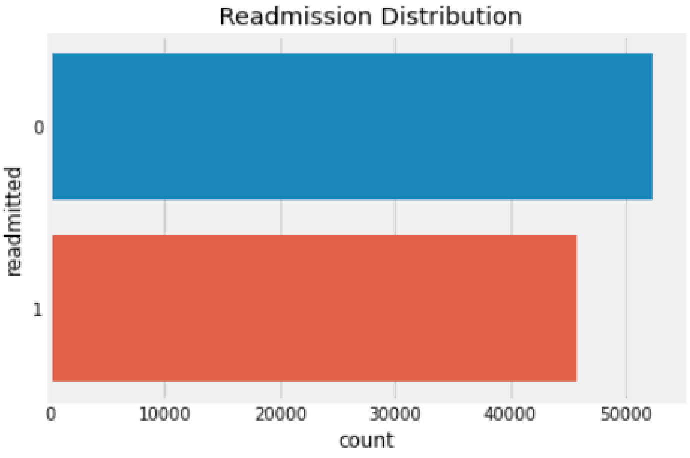


Figure 2: Readmission Distribution

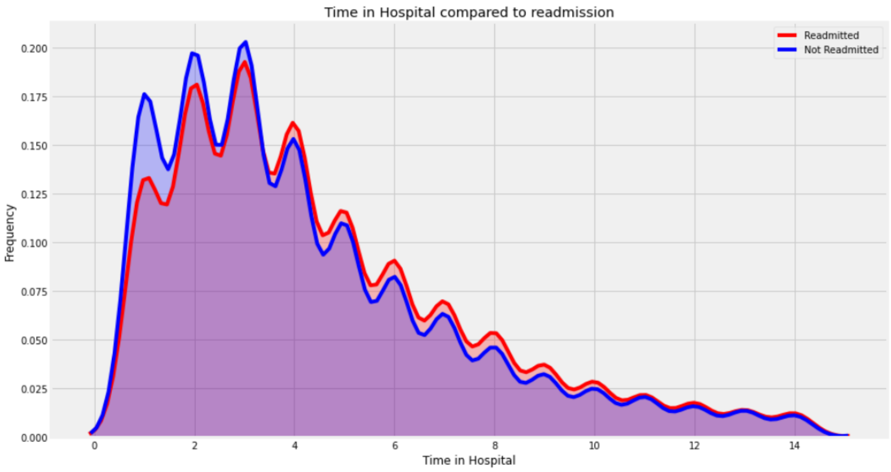


Figure 3: Time in Hospital compared to readmission

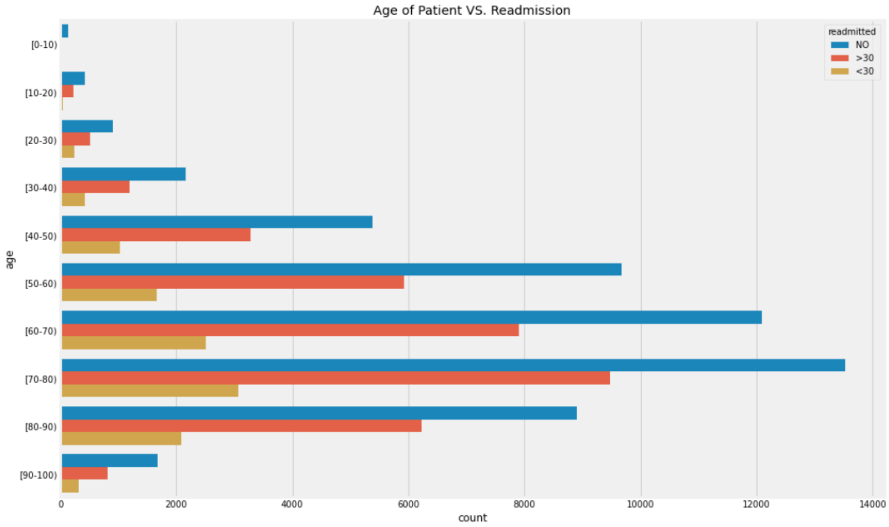


Figure 4: Age of Patient vs Readmission

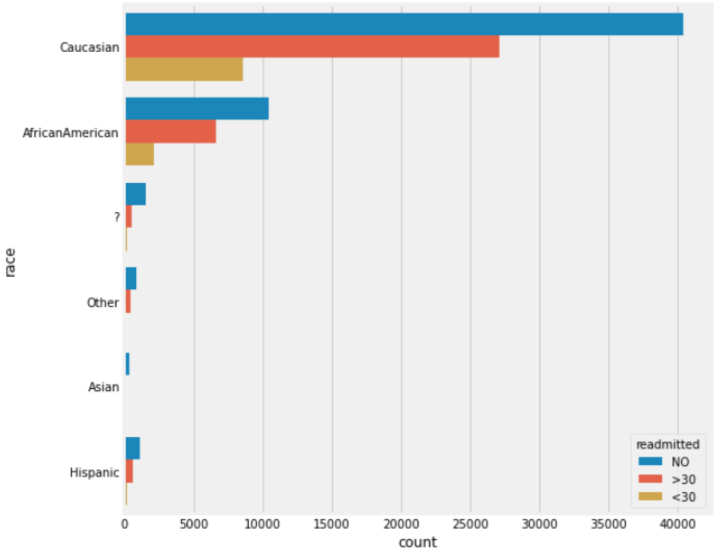


Figure 5: Race vs Readmitted

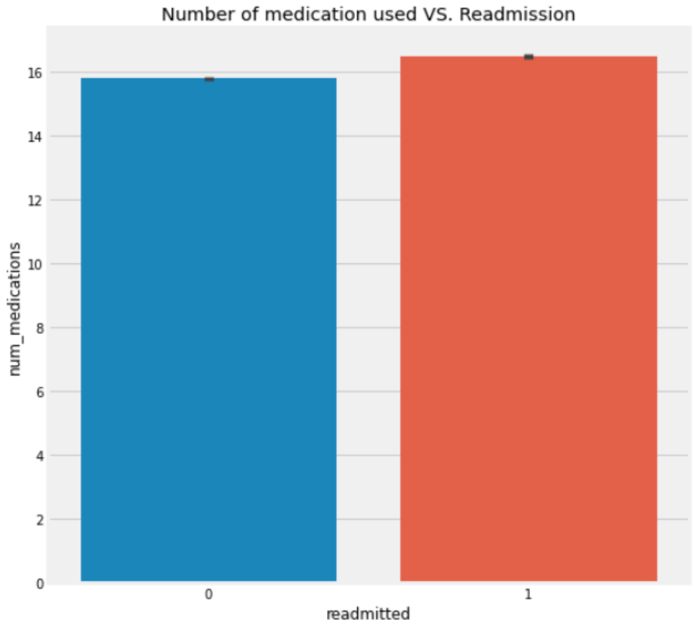


Figure 6: Number of medications used VS. Readmission

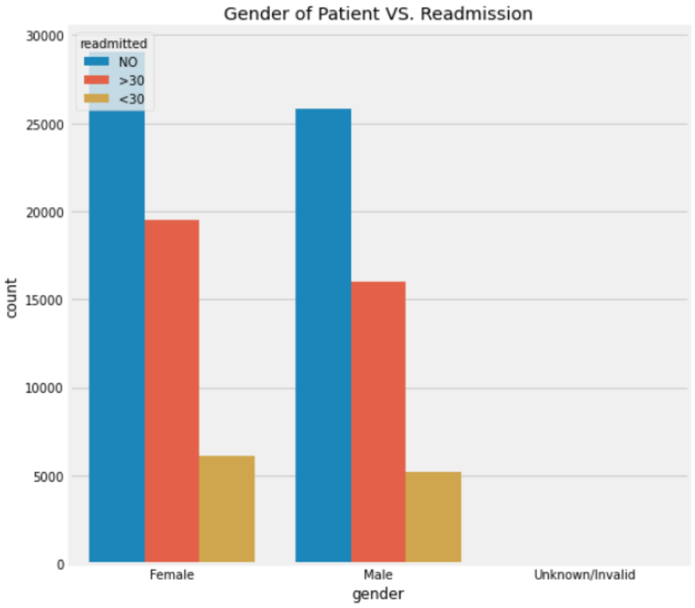


Figure 7: Gender of Patient VS. Readmission

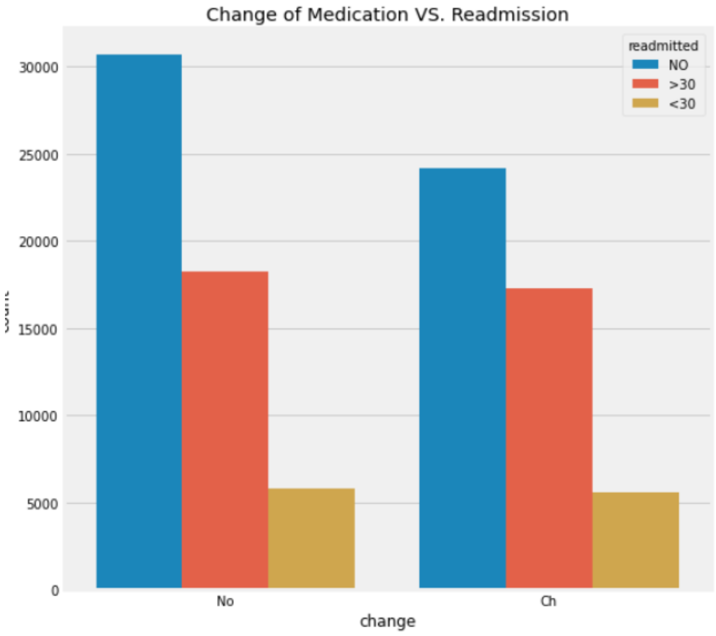


Figure 8: Change of Medication vs Readmission

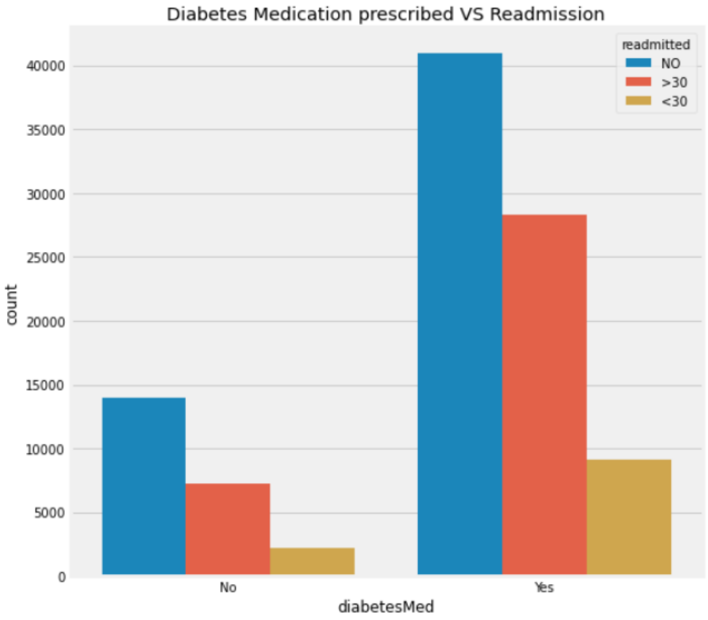


Figure 9: Diabetes Medication prescribed VS Readmission

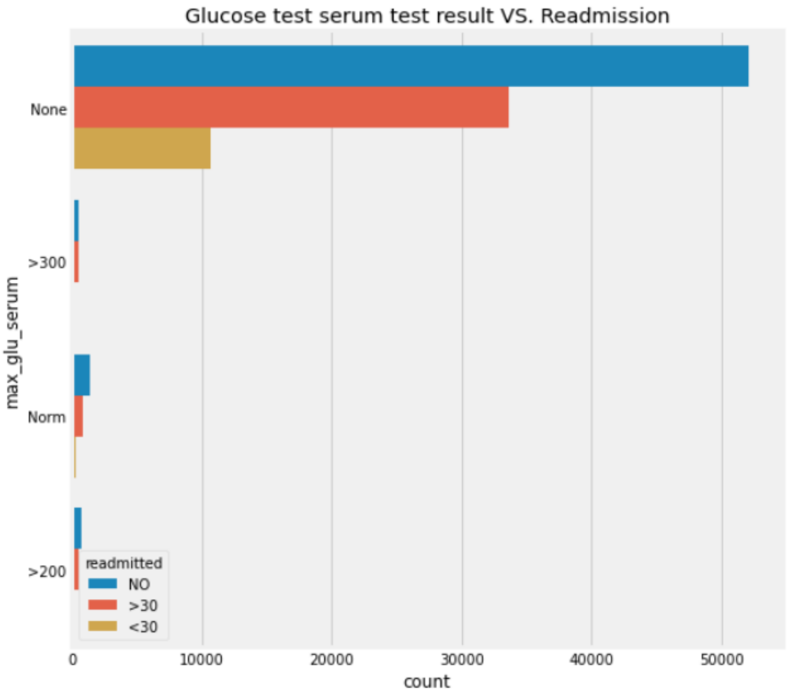
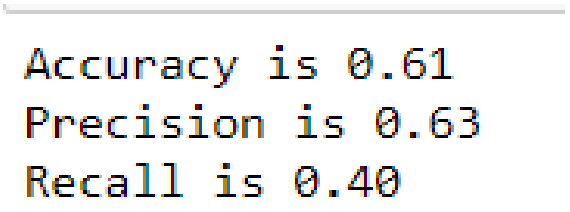


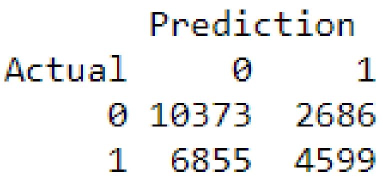
Figure 10: Glucose test serum test result VS. Readmission

Appendix B: Model Results and Statistics

Logistic Regression

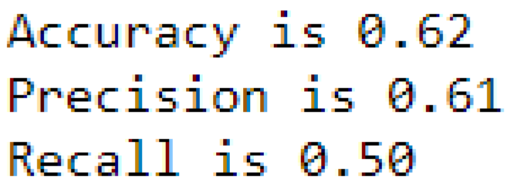


```
Accuracy is 0.61
Precision is 0.63
Recall is 0.40
```

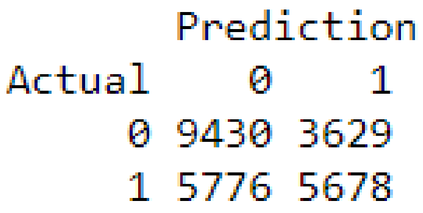


```
          Prediction
Actual      0      1
     0  10373   2686
     1   6855   4599
```

Decision Tree

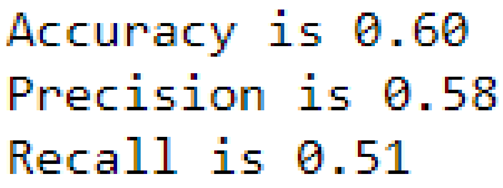


```
Accuracy is 0.62
Precision is 0.61
Recall is 0.50
```

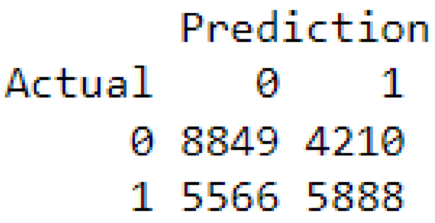


```
         Prediction
Actual     0    1
     0  9430 3629
     1  5776 5678
```

Random Forest

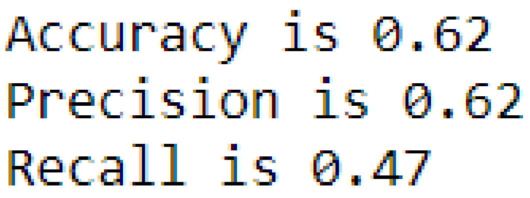


```
Accuracy is 0.60
Precision is 0.58
Recall is 0.51
```

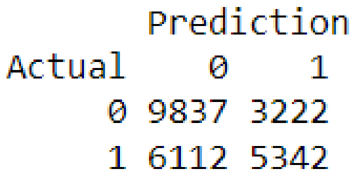


```
         Prediction
Actual     0    1
     0  8849 4210
     1  5566 5888
```

Gradient Boost

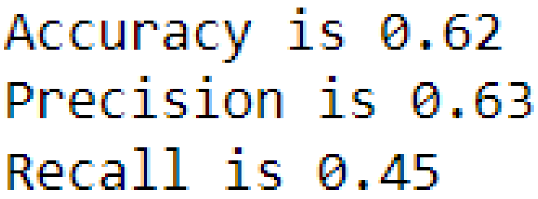


```
Accuracy is 0.62
Precision is 0.62
Recall is 0.47
```

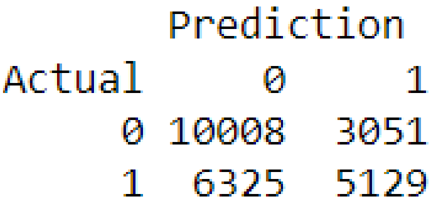


```
         Prediction
Actual     0    1
     0  9837 3222
     1  6112 5342
```

ADA Boost



```
Accuracy is 0.62
Precision is 0.63
Recall is 0.45
```



```
        Prediction
Actual     0     1
     0 10008  3051
     1  6325  5129
```