CISC451 Competition 2 Report

Group Member:
Steven Wen 20144322 Feiting Yang 20143750 Yifan Zhu 20146990

Package used:

Pandas: We use it to do fundamental data analysis. To use pandas, we just need to simply install using: pip install pandas, and then import pandas in python code.

Numpy: We use it to do mathematical calculations. To use it, we should use the ‘pip install numpy’ instruction and then import it into python code.

Sklearn: We use sklearn’s different package for data processing and modeling. This package can be installed by ‘pip install -U scikit-learn’.

KBinsDiscretizer: We use bin to group different categories into one. This is a package by Sklearn so we import it using ‘from sklearn.preprocessing import KBinsDiscretizer’.

Train_test_split: We use train test split for training the data. To import it we used ‘from sklearn.model_selection import train_test_split’.

Pipeline: We used a pipeline for the modeling to assemble more steps into one. To import this we used ‘from sklearn.pipeline import Pipeline’

RandomForestClassifier: We used random forest as our classifier. To implement this we used ‘from sklearn.ensemble import RandomForestClassifier’.

Data Preprocessing:

Missing Data:

We find that variable weight contains 96.8% of missing data and there is no good algorithm to fill in those, so we decided to drop those. Similar to payer_code and medical_specialty data that have 42.7% and 48.1% missing respectively.

For the other missing data like diag_1, diag_2,diag_3, gender and race since the missing value is less than 1% we decided to drop them. (Since KNN takes too ling to go though all the data)

Unique Value Data:

In this data set there are few values like 'citoglipton', 'examide' and ‘metformin-rosiglitazone’ only have one unique value which are not useful for the prediction, so we dropped it.

Categorical Data:

Then for some categorical data like ‘diag_1’, ‘diag_2’ and ‘diag_3’, we first transfer the value containing ‘V’ and ‘E’ into one category. Then use KBinsDiscretizer to group it in to 9 bins and all bins in each feature having the same number of points (quantile strategy) and return the bin identifier encoded as an integer value (ordinal encode).

Data Visualization:

To simply visualize the data. We draw the correlation graph as shown in figure1.

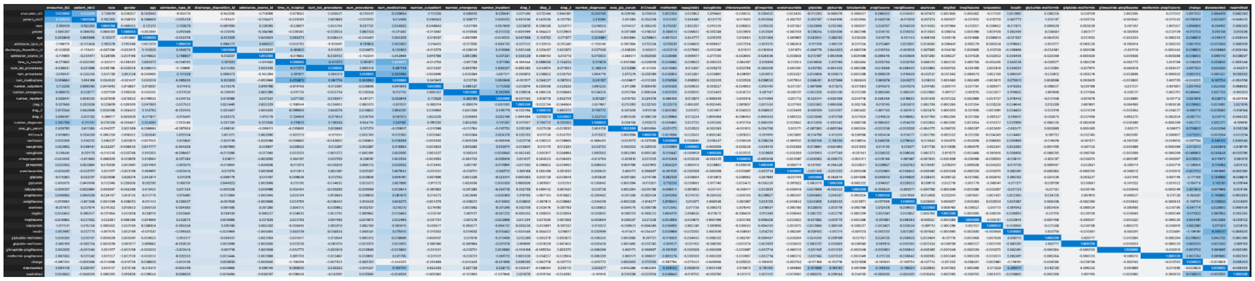


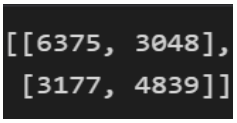
Figure1: Correlation of the variables

As the figure shows there is no correlation that is large enough for us to drop. Hence that the end of the data preprocess.

Model Implementation:

Two Label:

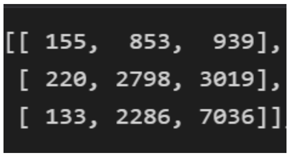
For whether the person will be readmitted or not we are using the random forest classifier. Since this is a classification problem and out of logistic regression, Navies Bayes and KNN. First of all, logistic regression can only solve linearly, and for part two of the question we need it to separate two different types so we did not consider using it. And for Navies Bayes, it needs independent features which some of the variables we have do have small correlation with each other. That's why we decided to use the decision tree model, but considering random forest leverages the power of multiple decision trees we decided to use it as our model. For this model our accuracy was: 64.3%, and the confusion matrix as:



```
[[6375, 3048],
 [3177, 4839]]
```

Three Label:

For three labels >30, < 30 and no. We first try to predict this data twice by predicting if the patient will be readmitted and if the date will be greater than 30 days. Later we found out that the train test is chosen as random but readmitted is encoded differently. The predictions are not using the same set of samples for the two predictions. We also found that the data set is unbalanced, so we used class_weight to balance the data, but this caused the accuracy dropped 1.5%(from 58.8%). For this model final accuracy is 57.3% confusion matrix is :



```
[[ 155,  853,  939],
 [ 220, 2798, 3019],
 [ 133, 2286, 7036]]
```