BUDT758B Final Project Report

Group 6 Team Supernova

Group Members: Haibo Yu, Yuchen Song, Quehe Feng,

Qian Jiang, Huanchen Fang, Lingfeng Wu

**Project Objective**

We aim to provide an accurate prediction of whether a customer will withdraw an account based on his/her related data through comparison of performances across models, including Logistic Regression, K-Nearest Neighbor, Keras Neural Network and so on.

**Data Source**

We retrieved our data from Kaggle and came up with our own business questions and solutions. Data can be found at <https://www.kaggle.com/santoshd3/bank-customers>

**Data Processing**

We made multiple modifications towards the original dataset in order for it to better serve our model training, where null values were dropped, unnecessary columns were removed and exploratory analysis was conducted. We generated multiple visualizations over different parts of the data to better illustrate and grasp the outline of our data. All the visualizations and tables we produced will be exhibited with markdowns explaining their purposes on the notebook we submit.

**Utilized Models**

**Logistic Regression:**

To solve a classification machine learning problem, the first model that came to our mind was logistic regression, where the target variable must be binary. In order to use this model, we created dummy variables for each categorical variable in the dataset, conducted a resampling due to imbalance of the target variable where 80% of the instances were 0. The accuracy of this model was 81%, which is not too ideal considering a 80% baseline accuracy. The inference over coefficients of our output suggests that age was the most decisive factor affecting customers’ decision where older customers are more likely to withdraw their account.

**KNN:**

KNN is very useful for unsupervised learning, so we choose KNN as a method for cluster analysis, which is used for exploratory data analysis to find hidden patterns or grouping in data.

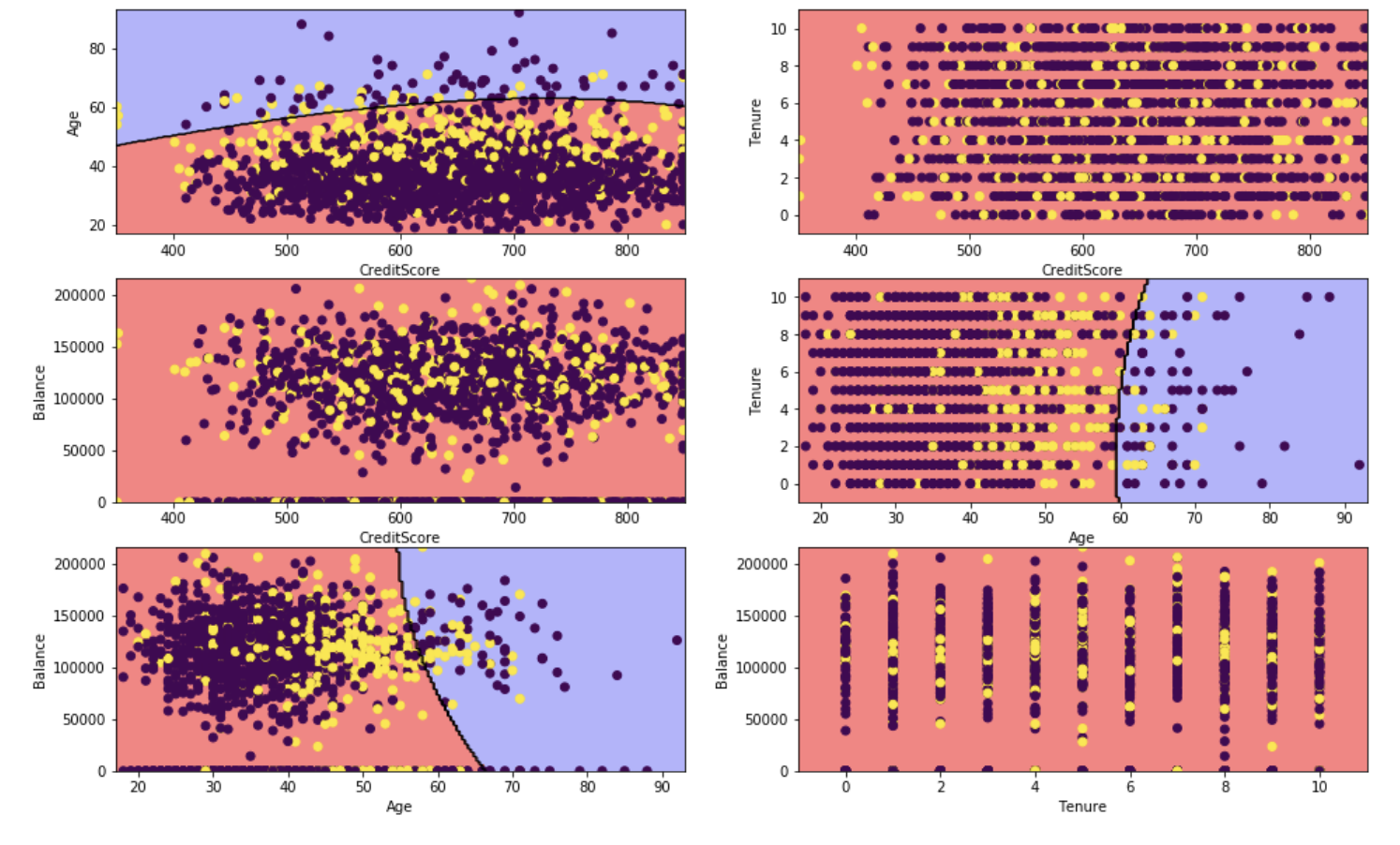
First of all we dug deeper into information about our data, for example, their distribution and relationships with the target variable. Starting with K= 5, we got our acc less than 80%. With a defined function to try 500 K values, we find the best k =28 with an 81% acc. We will definitely move forward to try more models, but what our insights and outputs for KNN model we can provide to our customer are:

* highly summarized customer attributes analysis chart for management level review
* How to segment customers into groups to provide the right service and monitor

**Quadratic Discriminant Analysis:**

Quadratic Discriminant Analysis is a kind of typical classifier, with, as its names suggest, a quadratic decision surface. In this QDA model, we assume that the group of customers who withdraw their account from the bank and the group of customers who do not withdraw their account have different Covariance Matrix. First Step in the QDA model, we proved that these two groups actually do have different covariances and so we can build up a QDA model for the bank. After changing types of several categorical features such as ‘Gender’ and ‘Geography’ from string to numeric, we train the model by splitting the test dataset and training dataset randomly.

The accuracy of the QDA model was 83.35%, which was much higher than our baseline model and the ROC is 0.620. So this QDA model is valid and can provide an insight for banks to explore features of customers that might lead them to withdraw their accounts. We could get the precision for Class 0 is 0.84 and for class 1 is 0.70, and the weighted average precision is 0.82 from the Classification Report and what we can conclude from the Report is that in this model, predicting class 0 is more accurate than predicting class 1. We suspect the reason is that this data is highly unbalanced for there are more 0 in the class, accounting for 79.6% of the total data. More data provided, better predictions.

Last but not least, the QDA works on our test dataset. In this bank customers dataset, QDA works well for Age vs Credit score, Age vs Balance as well as Age vs Tenure. As for other features, there does not appear to be a very clear quadratic decision surface. 

**Random Forest**:

After running the Random Forest Classification Model, We got an accuracy of 86.4%, which improved 6% from the baseline model. Then, we found the most important factor was age, which has the highest influence factor over whether customers will withdraw their account. Besides, estimated salary, credit score, balance and number of products were also important factors. In terms of geography, being a German makes one more probable to withdraw his or her account. Gender was the least important feature in the features we chose and males and females have similar performances.

**Sequential Deep Learning Model**

Sequential model is an artificial neural network model that can set multiple layers between the input and output layers. All the layers we can tune multiple hyperparameters to test performance of our model on this dataset. Some essential hyperparameters are activation function, units and initializer. The activation function is a function that calculates input data then output data to the next level. Units decide how many numbers of neurons to output to the next layer. Initializer define the way to set the initial random weights of Keras layers. In this case we start with “Relu” as activation function, 11 units and random normal distribution as initializer on our first layer. After experiments with different activation functions and initializers, we find out that the prediction accuracy stays around 84%. We also find out the best performance model is not the most complicated model.

**Linear Discriminant Analysis**

LinearDiscriminantAnalysis is a dimensionality reduction technique. It reduces the number of variables while retains the differences between the classes. After selecting features and converting the types of categorical features from string to numeric, I train the model. According to the coefficient numbers, Germany and olds has a higher probability to exit while active members and customers have high balance are less likely to exit. However, the model is unreliable: though the TNR is 96%, the AUC value is only 58% and the TPR is even only 20%. The whole accuracy is 81.3% while the baseline is 80.65%.

**Conclusion**

By analyzing all the models and their performance, we determine that Random Forest was the best model in predicting our target variable with 86% accuracy. In terms of inference, older people are more likely to withdraw their account. Geographically speaking, Germany has a higher withdrawal rate among all three countries in our dataset. Not surprisingly, credit scores and estimated salaries were also major players in affecting withdrawal rate.