# **Decision System of HELOC Application**

# **Introduction**

# In this report, I will design an interface containing a model to process information and help users. I will mention the whole process of what I did, including confirming my goals and assumptions, designing the model and doing performance evaluation, and the process of interface designing and making interpretation of the model.

**Goal and User Assumption**

**Goal:** The goal of the model is to predict whether the risk performance of the applicant is good or bad according to the features of the applicant. I also need to design an interface to get information about applicants and output my evaluation and conclusion for this applicant.

**Assumption of users:** Thetarget audiences are staff in banks who need to decide whether to provide mortgages to applicants. Therefore, I assume these users are people who understand complex concepts like “Percent Trades Never Delinquent” and can pre-calculate all the required ratios that I need. HoIver, they might not be very familiar with output coefficients related to statistics so I need to explain all these output results to them in a way that they can understand.

**Technique requirements:** Since I assume users do not have advanced skills in technique and statistics, I need to provide an easier interface to get information and show output. For my model, since I only have specific information and features to do the prediction, customers are required to only input the information and data that I asked them to provide. The extra information will not be able to process in model and the interface will only provide input boxes that require data that I required.

**Model design and Performance evaluation**

Firstly, I **processed the special values** in the HELOC data set. There are three types of special values: -7, -8, -9. I extended the data set with binary variables that indicate -7 and -8 in each of the columns and replaced them with the average value in each column. Besides, I drop the observations whose features have the value of -9.

Secondly, I want to **eliminate some unimportant features** to make the model more interpretable and improve the prediction accuracy. I tried three different ways: Recursive Feature Elimination (RFE), Correlation, and Random Forest. When I applied the **Recursive Feature Elimination (RFE)** method to select the features, I created a for loop to iterate the number of features from 25 to 34. And then run the linear models and tree models, including Logistic Regress, SVC, Naive Bayes, Decision Tree, Random Forest, and Boosting models. I found LDA models with more than 30 features perform better than other linear models and tree models. The results are shown below in Fig.1. The correlation way and Random Forest way also gave us the same indication.

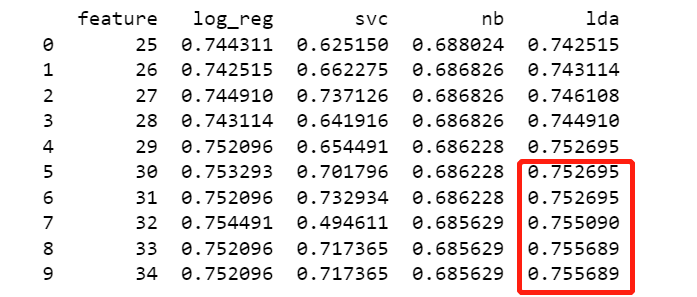


Figure 1 Linear Model Accuracy

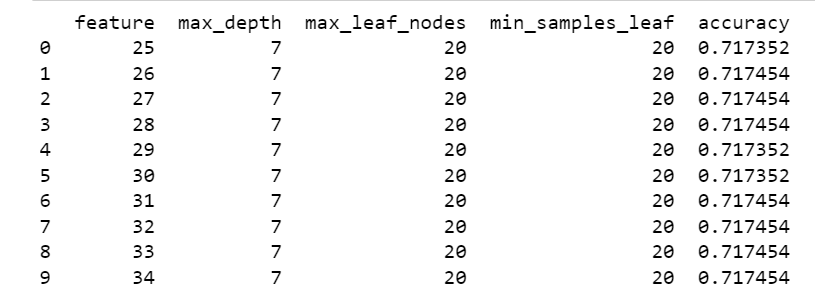


Figure 2 Decision Tree Model Accuracy

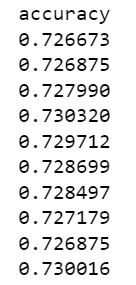
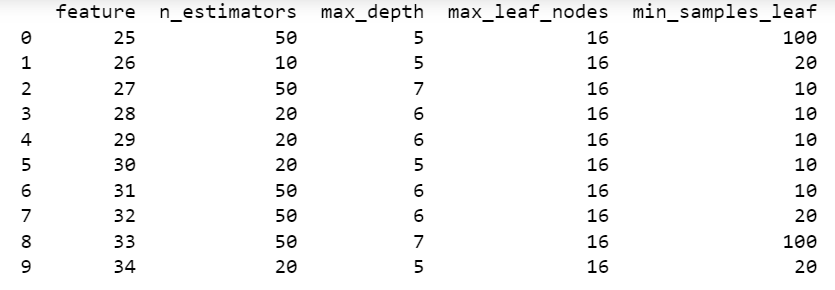


Figure 3 Random Forest Model Accuracy

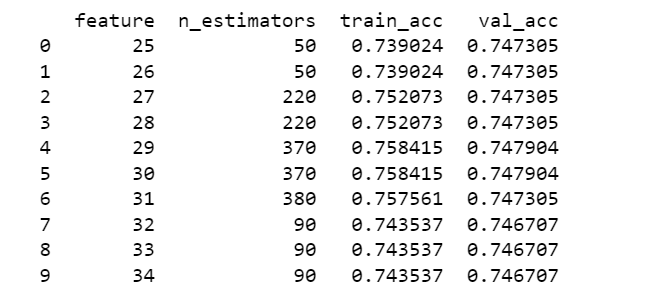


Figure 4 Boosting Model Accuracy

HoIver, accuracy is not the only criteria I care about. As different types of false predictions have different costs, I need to trade off these types. According to my assumption, type I error is taking good risk performance as a bad one; Type II error is taking bad risk performance as a good one. And the cost of Type II error is largely higher than type I error. As a result, I hope Type II errors are as few as possible and **FOR** will be a good indicator. **FOR** is the probability that I predict a customer's risk performance will be good, but actually the risk performance is bad. Therefore, I calculatedand compared the **FOR** of all LDA models with more than 30 features. The result is shown in Fig.5. The LDA model with 33 features performs best.

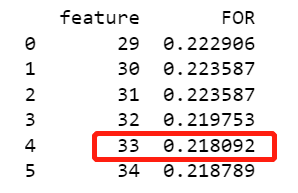


Figure 5 FOR Performance

Finally, **I choose the LDA model with 33 features as final model** and its features are shown in Fig.6. Then I used cross-validation to check the model's accuracy again. The average accuracy is 0.7325 and the standard deviation is 0.0045. The 95% confidence interval is [0.7237, 0.7415].

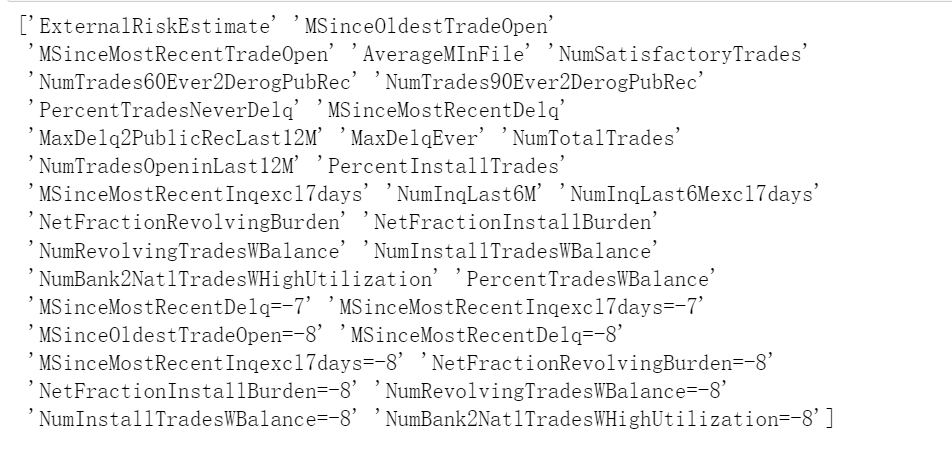
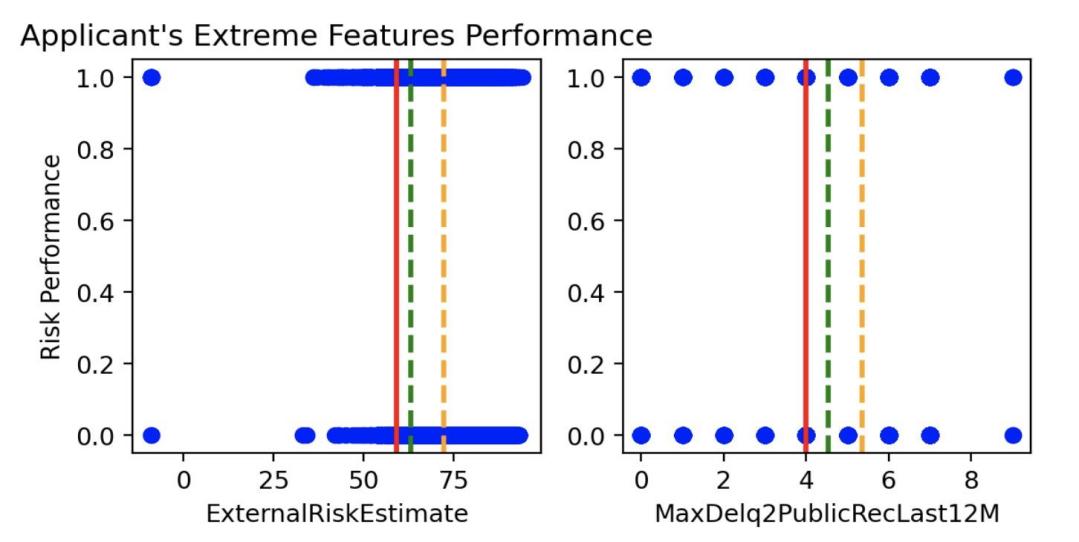


Figure 6 Features Selection

**Interface and Prediction Interpretation**

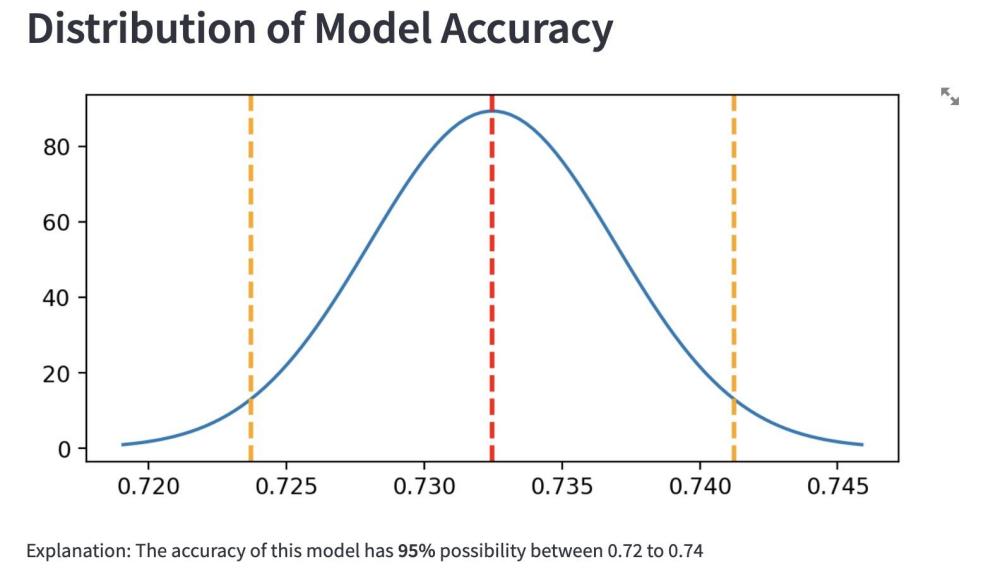
**Interface:** The first part of the interface is an APP introduction, which describes this app’s function, users, and using requirements in text. The Model requires that the user has the Bureau Record and Investigation information of the applicant, which is due to the deletion of observations containing -9 in the LDA model. Next part is the applicant's information. The user needs to input 23 features of the applicant in the number input areas and then get the model prediction results.

**Prediction Interpretation:** To help the user explain the prediction to the applicant, I decided to show features that have very large differences from their average values of high risk applicants and low risk applicants computed from the HELOC dataset. I present values of these 6 ‘extreme’ features and the corresponding average level of the feature using a scatter with 3 lines, which helps the user understand the performance of the applicants more easily. To give an example(This is 2 of 6 features for 1 applicant whose application should be refused):



The feature value of this applicant(red line) is closer to the feature average value of the refused application(green line) than the feature average value of the accepted application(yellow line), which indicates the bad performance of these 2 features of the applicant and also explains the reason why this particular applicant should be refused.

**Distribution of Accuracy:** I show the accuracy confidence interval and distribution of the model in several cross validation to prove that the model is reliable and accurate in prediction.



# **Conclusion**

# In conclusion, I designed an interface with a model that can help bank staff decide whether to provide mortgages to applicants or not. In the model selection process, I used RFE, correlation, and Random forest to find features that I want to use in the model. After comparing different models’ accuracy and FOR after applying these three methods, I found that **LDA with 33 features is the model** that performed best. Next, I designed the interface. The user will see the extreme feature and final decision after the input feature of the applicant.