**CIS 5570 - Introduction to Big Data Final Project**

**German Credit Risk Prediction**

Team Members:

Wenqing Shi

Yuting Wang

Ying Zhao

**1.Abstract**

A credit risk is the risk of default on a debt that may arise from a borrower failing to make required payments. The classification is a major fundamental method for evaluating the credit risk in our project. In this study, we use Python 3.0 to predict each person as good or bad credit risks according to the age, sex, job, housing status, saving accounts, checking account, and credit amount. After preparing the data, we use logistic model and random forecast to train data, and then apply test set to get results. And consequently, using learning curve (or other evaluation method, ROC, precision, recall rate, accuracy) to evaluate the prediction, it helps us make some adjustments of parameters.

**2.Introduction**

The commercial banks are high-indebted and high-risk industries based on credit and operation in currency lending and settlement. Therefore, it is crucial for a bank to judge a person’s credit quality. In our project, our dataset contains 1000 entries with 20 categorial/symbolic attributes prepared by Prof. Hofmann. In this dataset, each entry represents a person who takes a credit by a bank. Each person is classified as good or bad credit risks according to the set of attributes.

**3.Related work**

One of the most famous projects on Kaggle is prediction of survival on Titanic. It needs to analysis the different elements of luck involved in surviving the sinking, such as women, children, and the upper-class, by applying the tools of machine learning to predict which passengers survived the tragedy.

In that project, firstly, use binary classification to classify samples, then analysis the correlation of different classes with the solution goal; Secondly, prepare the data, which includes cleansing and making up missing data, converting text values into numeric values by generating dummy variables; Thirdly, approach the problem by find out which features contribute to the solution within the training dataset; Lastly, build the prediction by logistic model(or other algorithms) to predict the goal on test dataset and evaluate the results by learning curve, AUC or ROC curve .

**4.Methodology**

**4.1 Algorithm Approach: Binary classification and Logistic Regression Model**

Binary classification is the task of classifying the elements of a given set into two groups (predicting which group each one belongs to) based on a classification rule.

Logistic model is a widely used statistical model that, in its basic form, uses a logistic function to model a binary dependent variable; In binary logistic regression, the outcome is usually coded as "0" or "1". If an observed outcome for the dependent variable is possible, it is usually coded as "1" and the contrary outcome as "0". The goal is to model the probability of a random variable Y being 0 or 1 given experimental data.

### Logistic model:

Y= a + b1x1 + b2x2 + … + bkxk

Probability =

**4.2 Reason to choose:**

Logistic regression is used in various fields, including machine learning, most medical fields, and social sciences. In economics, it can be used to predict the likelihood of a homeowner defaulting on a mortgage, so it is a suitable model to predict whether a person will have good or bad credit risks according to the set of attributes.

**4.3 How to apply:**

After learning from Titanic project and make some research on logistic model, firstly, we need to prepare the data, cleanse and make up the dataset, acquire training and testing data ;Secondly, use binary classification to classify samples, analyze, identify patterns, and explore the data then analysis the correlation of different classes with the solution goal; Then the most important part is to model, predict and solve the problem by logistic model; Lastly, we need to evaluate the results by ROC curve.

**5.Experimental Discussion**

**5.1 Data Preparation**

1) Acquire training and testing data: We use 700 users’ data as training data and 300 users data as testing data.

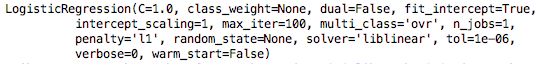
2) Basic acknowledge: By using Python script, we can get the description of data, including looking the Type of Data, null numbers, unique values and the first rows of our dataset; then use bar charts and scatter plots to know the column and distribution of different features. There are 10 IVs (like the following) and 1 DV, which is Risk.

Age (numeric)/Sex (text: male, female)/Job (numeric: 0 - unskilled and non-resident, 1 - unskilled and resident, 2 - skilled, 3 - highly skilled)/Housing (text: own, rent, or free)/Saving accounts (text - little, moderate, quite rich, rich)/Checking account (numeric, in DM - Deutsch Mark)/Credit amount (numeric, in DM)/Duration (numeric, in month)Purpose (text: car, furniture/equipment, radio/TV, domestic appliances, repairs, education, business, vacation/others)

3) Explorations: After plot correlation matrix, we do not find multicollinearity; We use mean value and new categories to fulfill the null values and convert ‘age’ into categorical variable, use standard normalization to deal with ’credit amount’, and generate dummy variables.

* 1. **Model, predict and solve the problem**

After using 3 logistic models and 1 random forest model, comparing the metrics of the model results, we get the best model in condition of: 1. Make up null value with mean value; 2. Convert Age into 4 classes. 3.Use standard normalization to deal with credit amount; 4. Keep all 10 features;5. Use logistic model.

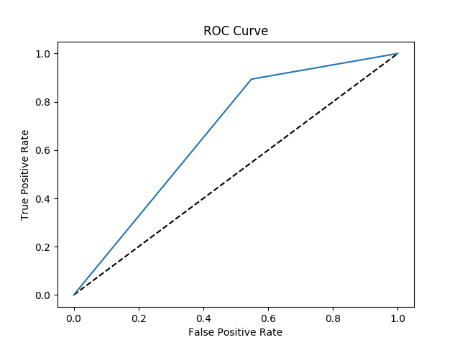
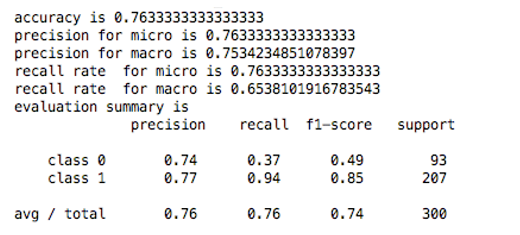


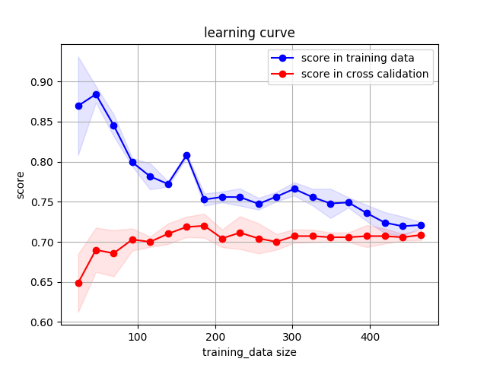
We use the best model to predict the test data and get the prediction showed in appendix.

* 1. **Evaluation**

We can find that the accuracy of Model1, 2, 3 and random forest are 0.743,0.746, 0.763 and 0.7566 respectively. After compared accuracy, we can see that **Model 3** is the best model,

Model3 Accuracy:



From the Roc Curve and learning, we can know that the AUC value= 0.763. If AUC is larger than 0.7, it proves that the model works well. The learning curve of blue line and red line are very close to each other, so the result is great.

**6.Contribution**

In our project, we divide our work into three parts.

For the team member Wenqing Shi, she contributes in technology part, such as write code, provide results and plots materials, and then revise and help team members finish the project report.

For the team member Yuting Wang, she finds the project theme and dataset, make some research on logistic model, Titanic project and model evaluation, and then write project reports and make project power point.

For the team member Ying Zhao, she prepared the background information, making the power point, organize the document, write the report and searched related information.

**7.Conclusion**.

1)Comparing the results from application of Model 3 on test dataset, we can see that Model 3 works well.

2)The customers who rent a house has higher probability to make a risk; Saving accounts and checking accounts are little or moderate are more likely to face risk; The larger Credit amounts are, the higher the risks are; The younger are the customer, the higher of risk they will have to face the risk.

3)In the future, we need to pay attention to the multi-collinearity between credit amount and duration. We can add more training data, involve more features and try more models.

4) Checking accounts have 40% missing values in total, which may have misleading to our model. We need to consider other features to replace it.

**8.References**

[1] Basel Committee on Banking Supervision. BIS. September 2000. Retrieved 13 December 2013.

[2] <https://www.kaggle.com/kabure/predicting-credit-risk-model-pipeline/data>

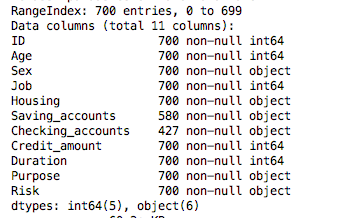
[3] https://www.kaggle.com/startupsci/titanic-data-science-solutions

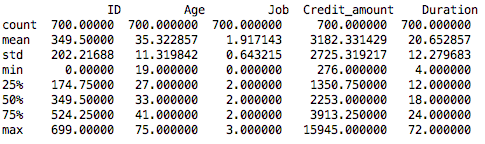
[4] <https://en.wikipedia.org/wiki/Logistic_regression>

[5] <https://en.wikipedia.org/wiki/Binary_classification>

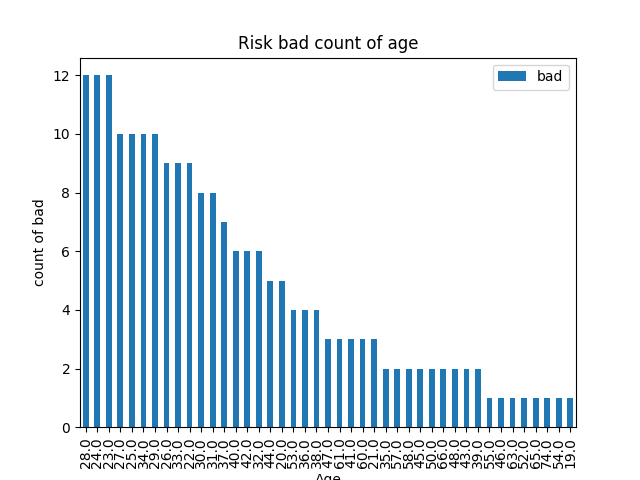
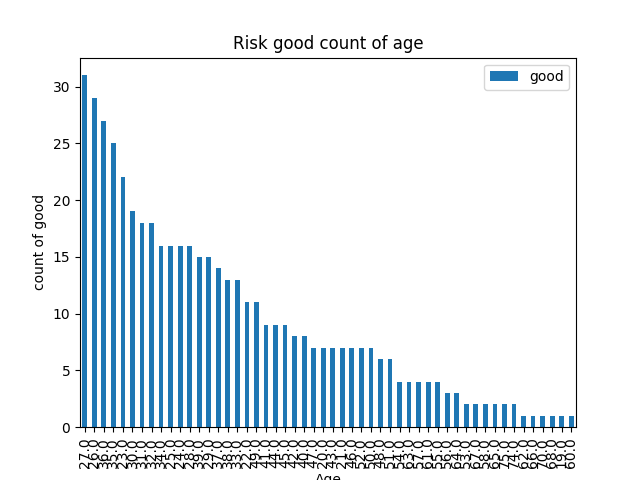
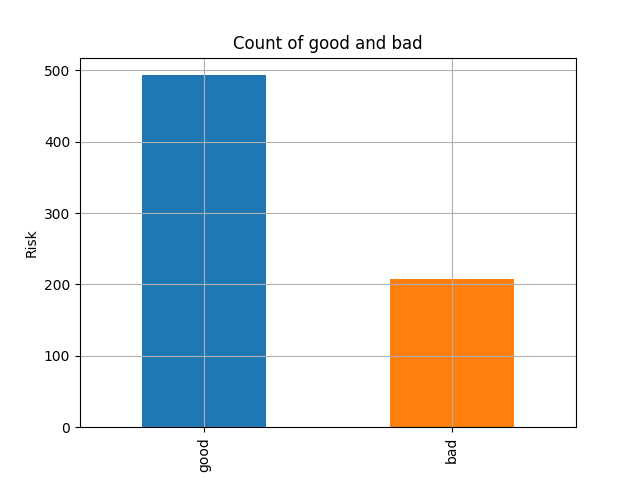
**Appendix**

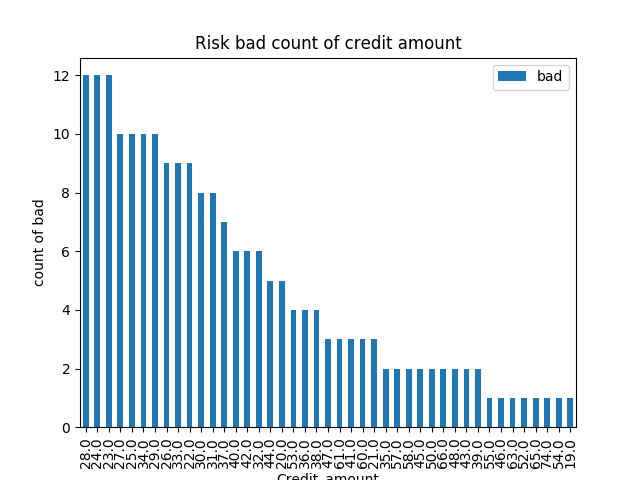
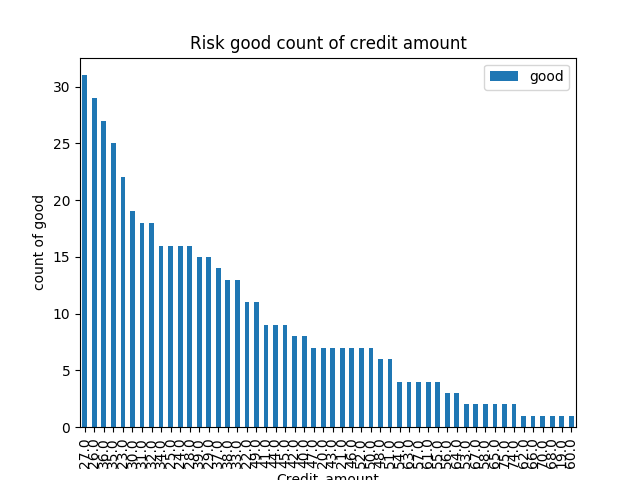
1. data description

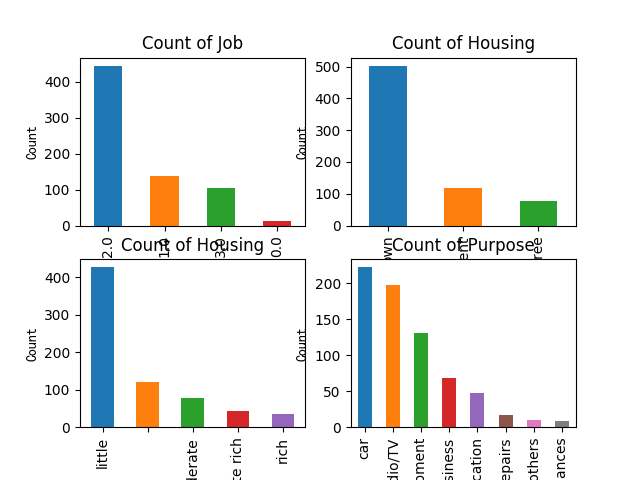
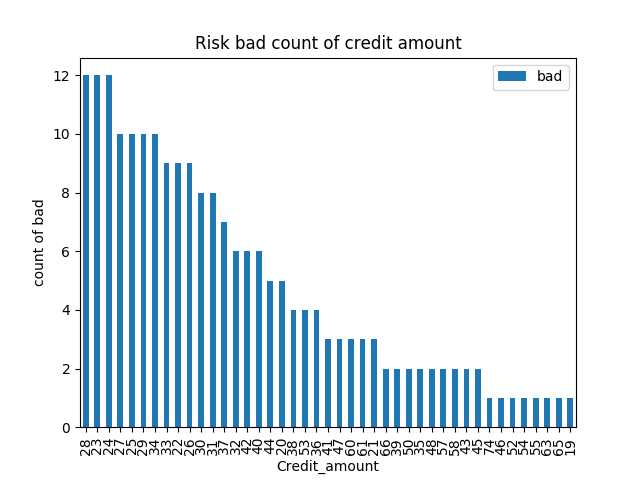
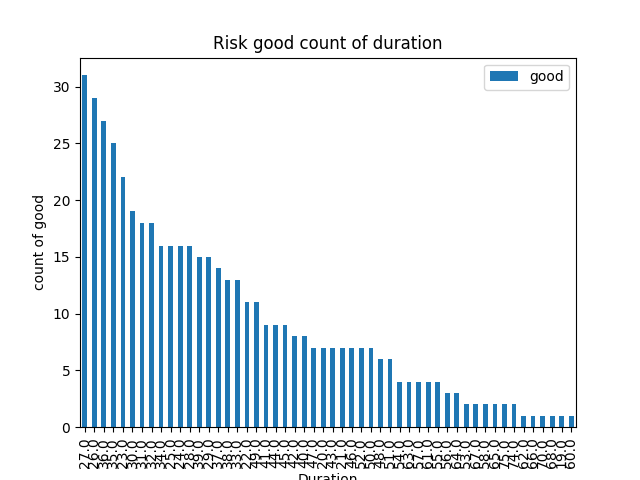


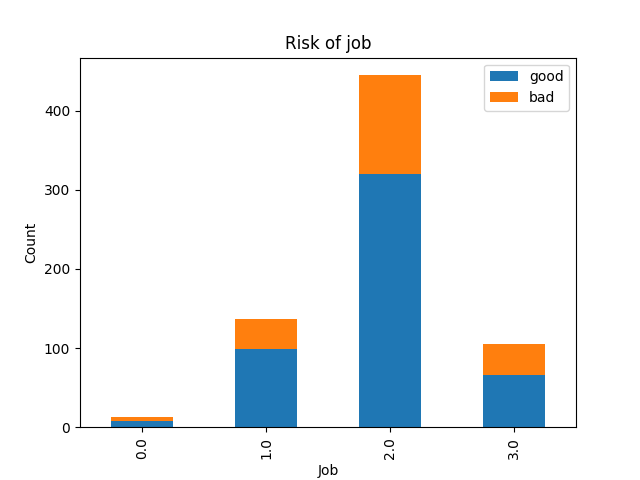
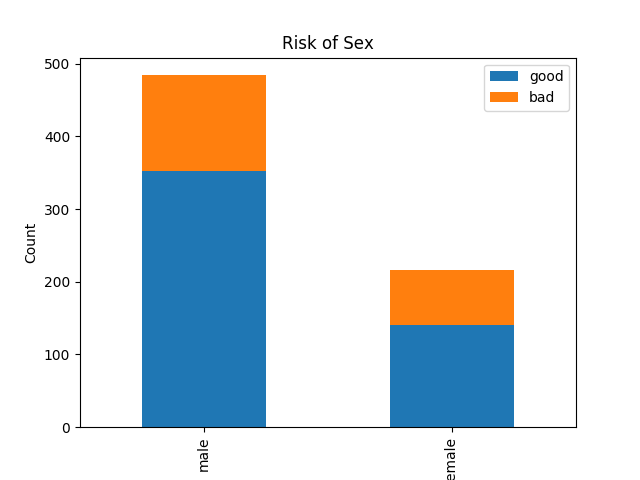


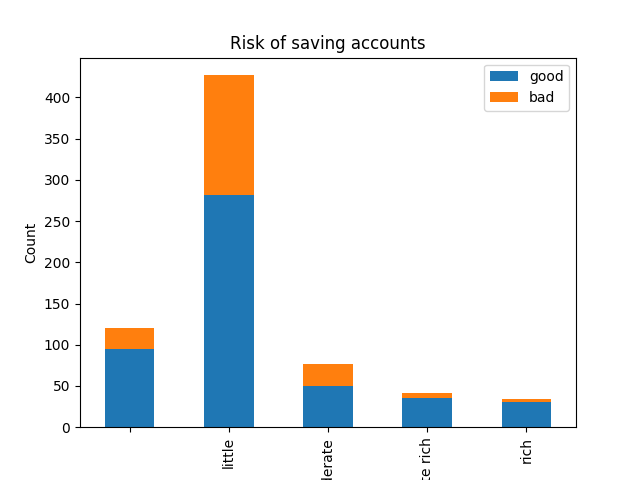
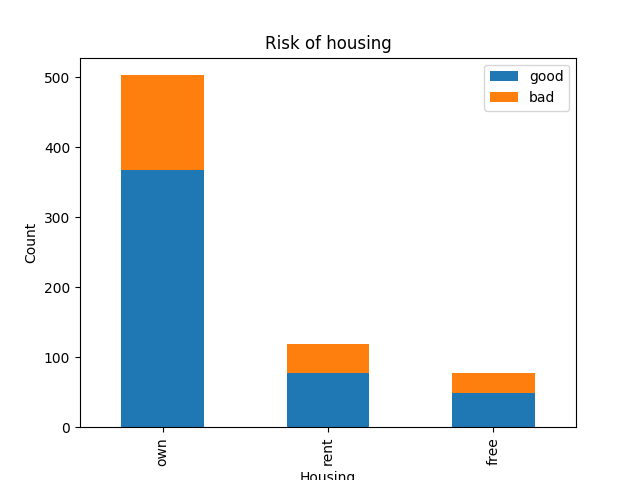
2. plots(all features are valid)

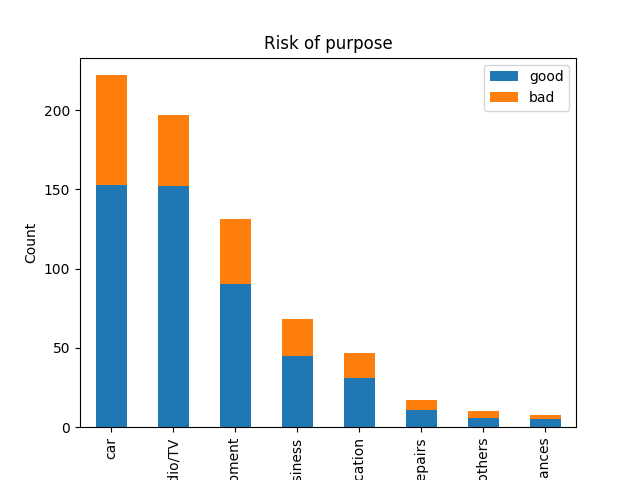
c



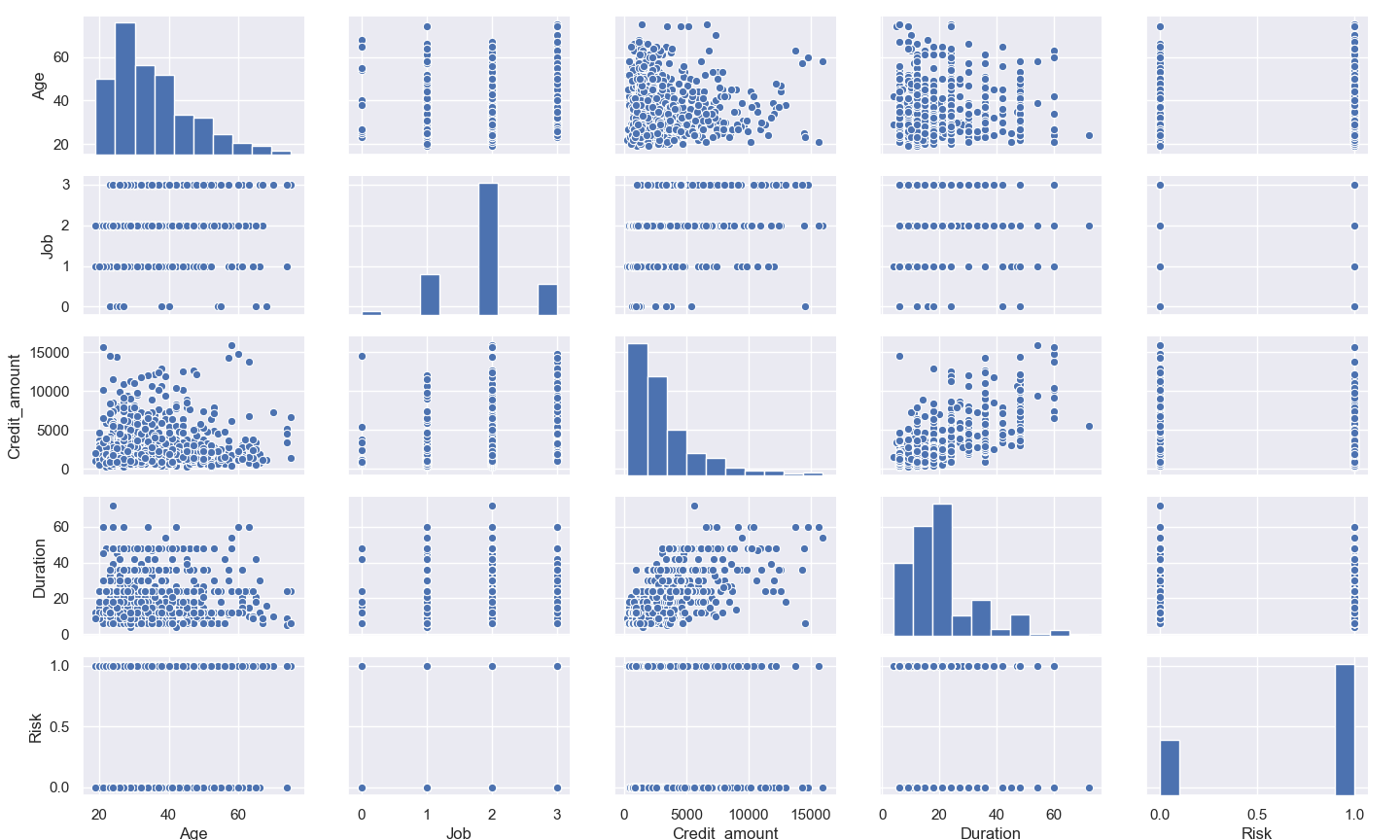








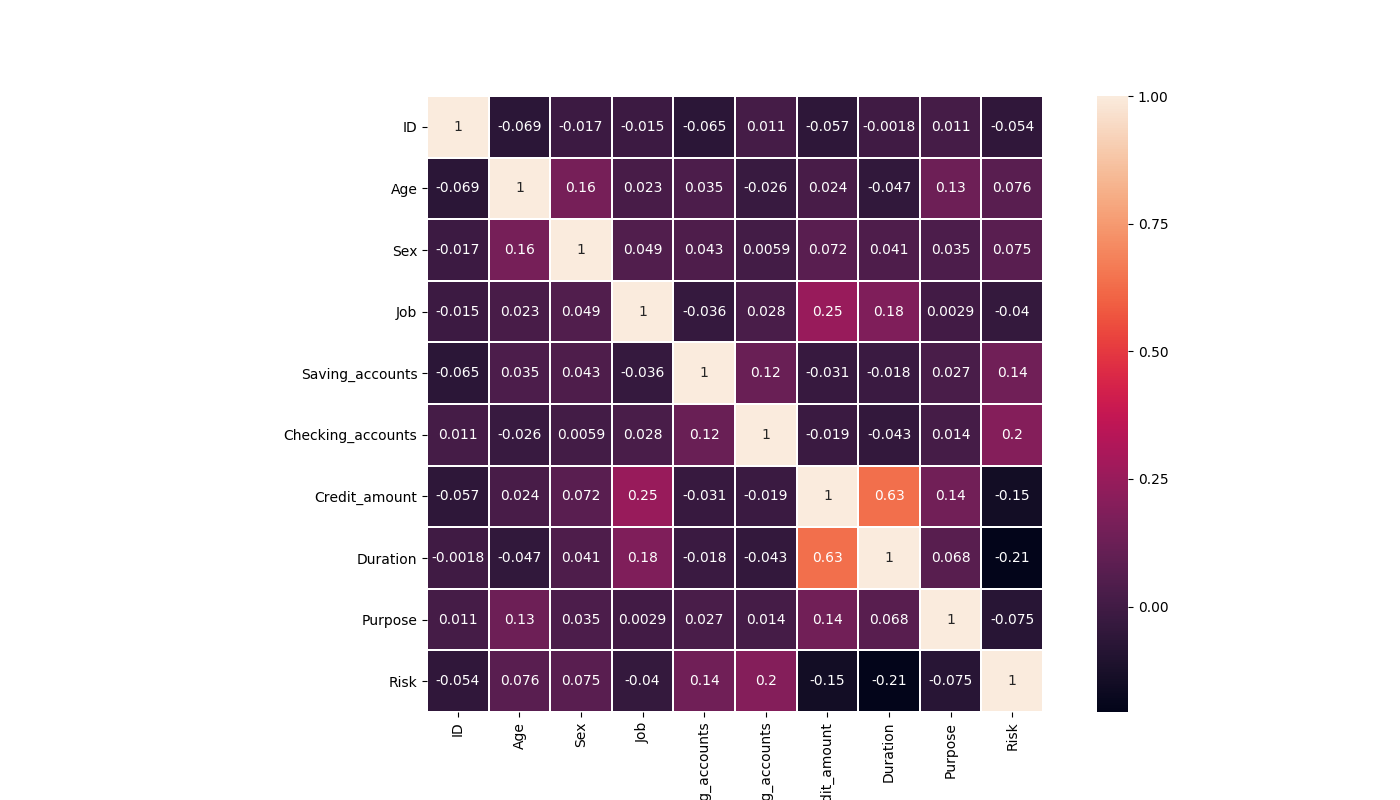
scatter plots



(plots to be added)

3. prepare data

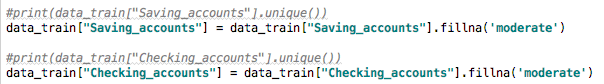
plot correlation matrix, not find multicollinearity (model 1)



deal with missing data, add a new category “no\_info” (model 1)



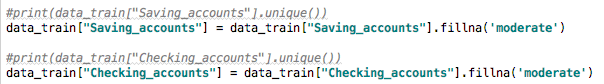
deal with missing data, add mean values (model 2)



convert age to categorical variables (model 2)



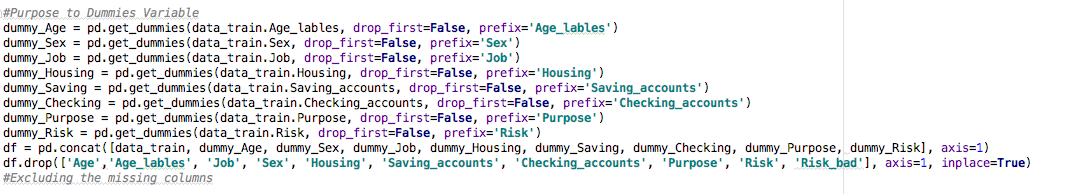
deal with missing data, add mean values (model 3)



convert age to smaller categories, compared to model 2(model 3)

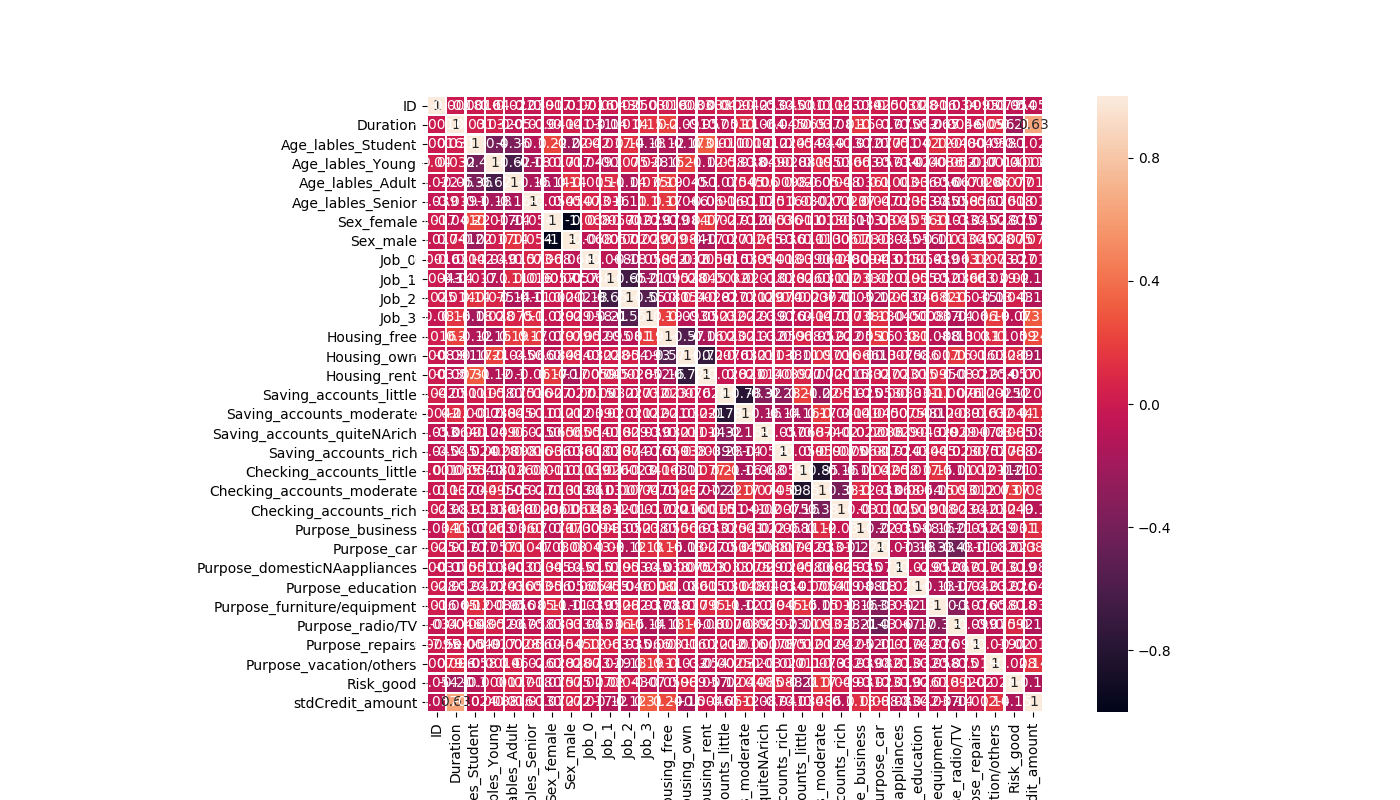


get dummy variables



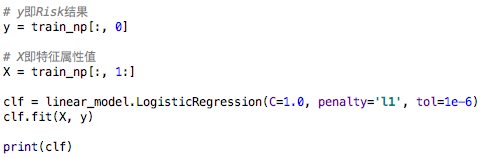
plot correlation matrix, still not find multicollinearity



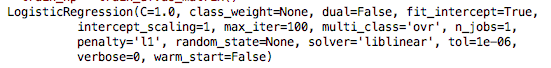


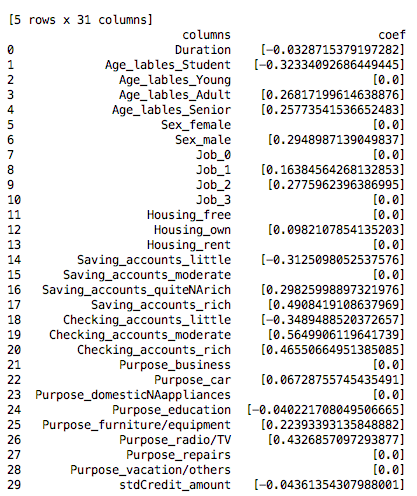
4. using logistic regression model

clf is the model

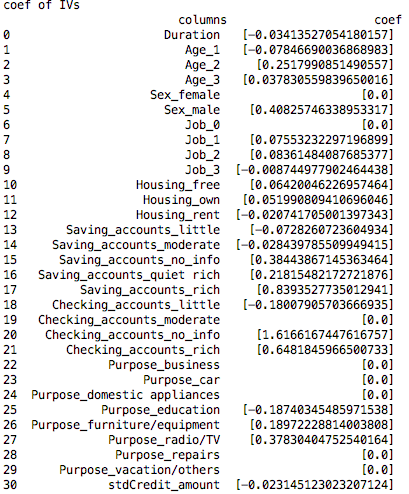


model 2





Model 3(best model)

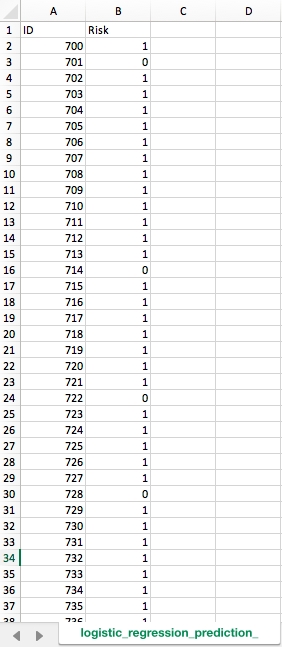


5. deal same with test data

6. get the prediction

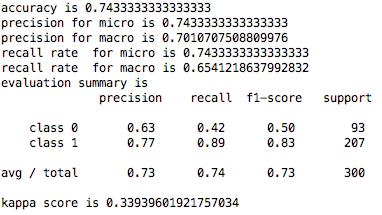


7.Application:

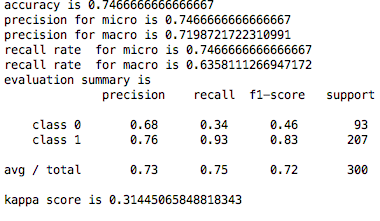


8. evaluation result & 8. Improve model

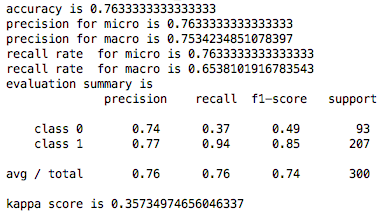
model 1



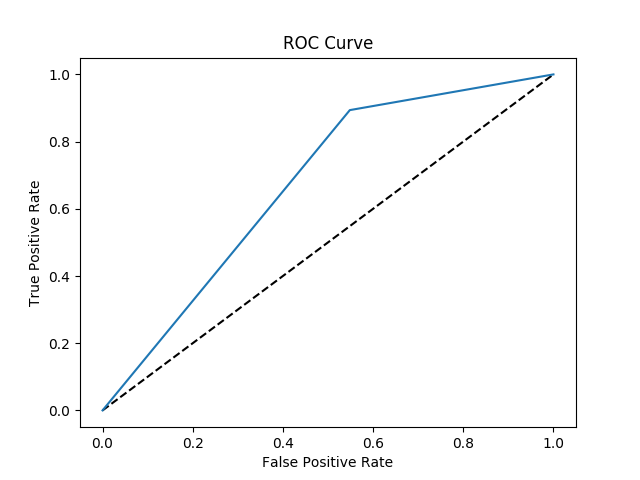
model 2



model 3(best model)

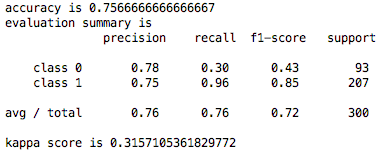


mode 3 ROC curve(best model)



Another method with model 3 data(random forecast)

not better than mode 3, so mode 3 is our best model



random forecast roc curve

