

Ran Bi  
CAPP30254

## **Assignment 2: Machine Learning Pipeline**

Credit scoring is critical to commercial bank's loan decisions. In this assignment, I built a preliminary machine learning pipeline to analyze a dataset containing 150,000 individual's personal and credit information. I built three classifiers to predict whether somebody would experience financial distress in the next two years, and evaluated them based on three metrics.

### **1. Read and explore data.**

The descriptive statistics summary [Table 1] and distribution graphs indicate the dataset may not be a representative sample of borrowers: only 6% of the sample experienced 90 days past due delinquency or worse. The average age (over 50), monthly income (6670) and number of dependents (0.75) suggest that individuals in this sample are relatively middle-aged, well-off and live in a small household without many dependents.

The correlation table [Table 2] suggests that number of times borrower has been 30-59 days past due but no worse in the last 2 years is mostly associated with delinquency in the next two years.

### **2. Imputation and feature engineering**

The dataset suffers from missing value problem. 19.82% of monthly income and 2.6% of number of dependents data are missing. I imputed number\_of\_dependents by 0 – I assume individual did not have dependents if he or she did not specify the number of dependents. I imputed monthly income by sample mean. For future reference I may impute by KNN.

I discretized age variable into 14 bins: (0,20), (20,25), (25,30)...(75,80), (80,110). Note that I created larger bins near the min/max values to account for outliers. I also discretized debt ratio by Quantile, and then converted such categorical data into dummy variables.

### **3. Build and evaluate classifiers.**

I split data into training set and testing set. I used the training set to train three classifiers where all features were considered as predictors: logistic regression model, K-Neighbors model, and decision tree model. Then I used the testing set to make prediction for each model, and evaluated model based on three metrics: accuracy, recall and precision. All three models achieve accuracy score of around 93%. However, bearing in mind that 93% of the individuals in the sample were labelled as False, the three models do not exhibit much predictive power. The fairly low accuracy scores and recall scores further echo this point.

*Selected tables and graphs are in the appendix. All files and graphs are saved in output folder.*

Variables	count	mean	std	min	25%	50%	75%	max	missing value count
person_id	150000	75000.5	43301.41453	1	37500.75	75000.5	112500.25	150000	0
serious_dliqin2yrs	150000	0.06684	0.249745531	0	0	0	0	1	0
living_utilization_of_unsecured_lines	150000	6.048438055	249.7553706	0	0.029867442	0.154180737	0.559046248	50708	0
age	150000	52.29520667	14.77186586	0	41	52	63	109	0
zipcode	150000	60648.81001	56.74819728	60601	60625	60629	60644	60804	0
ne30-59_days_past_due_not_worse	150000	0.421033333	4.192781272	0	0	0	0	98	0
debt_ratio	150000	353.0050758	2037.818523	0	0.175073832	0.366507841	0.868253773	329664	0
monthly_income	120269	6670.221237	14384.67422	0				3008750	29731
er_of_open_credit_lines_and_loans	150000	8.45276	5.14595099	0	5	8	11	58	0
number_of_times90_days_late	150000	0.265973333	4.169303788	0	0	0	0	98	0
number_real_estate_loans_or_lines	150000	1.01824	1.129770985	0	0	1	2	54	0
ne60-89_days_past_due_not_worse	150000	0.240386667	4.155179421	0	0	0	0	98	0
number_of_dependents	146076	0.7572222268	1.115086071	0				20	3924

Table 1. Descriptive Statistics

**Table 2. Correlations**

Variables	serious_dlqin2yrs
revolving utilization of unsecured lines	-0.001801503
age	-0.115385518
zipcode	0.005103214
number_of_time30-59 days past due not worse	0.125586965
debt_ratio	-0.00760212
monthly_income	-0.019745547
number of open credit lines and loans	-0.029668568
number_of_times90 days late	0.117174613
number real estate loans or lines	-0.007038116
number_of_time60-89 days past due not worse	0.102260861
number_of_dependents	0.046047944

**Table 3. Classifier Evaluation**

LogisticRegression

Accuracy score is: 0.93464

Recall score is: 0.04388459975619667

Precision score is: 0.5242718446601942

KNeighborsClassifier

Accuracy score is: 0.9329066666666667

Recall score is: 0.025599349857781388

Precision score is: 0.34806629834254144

DecisionTreeClassifier

Accuracy score is: 0.9

Recall score is: 0.28199918732222673

Precision score is: 0.25924542398206946





