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 perdictors: 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.

146076 0.757222268	ne60-89 days past due not worse 150000 0.240386667 4.155179421 0	number_real_estate_loans_or_lines	number_of_times90_days_late	er_of_open_credit_lines_and_loans	monthly_income 120269 6670.221237 14384.67422 0	debt_ratio	ne30-59_days_past_due_not_worse	zipcode 150000 60648.81001 56.74819728 60601	age 150000 52.29520667 14.77186586 0	ving_utilization_of_unsecured_lines 150000 6.048438055 249.7553706 0 0.0	serious_dlqin2yrs	person_id 150000 75000.5 43301.41453 1	Variables count mean std min	
	0	0	0	₅		175073832	0	60625	41)29867442	0	37500.75	25%	
	0	ר	0	8		0.366507841	0	60629	52	0.154180737	0	75000.5	50%	
	0	2	0	11		0.868253773	0	60644	63	0.559046248	0	112500.25	75%	
20	98	54	98	58	3008750	329664	98	60804	109	50708	1	150000	max	
3924	0	0	0	0	29731	0	0	0	0	0	0	0	count	value

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

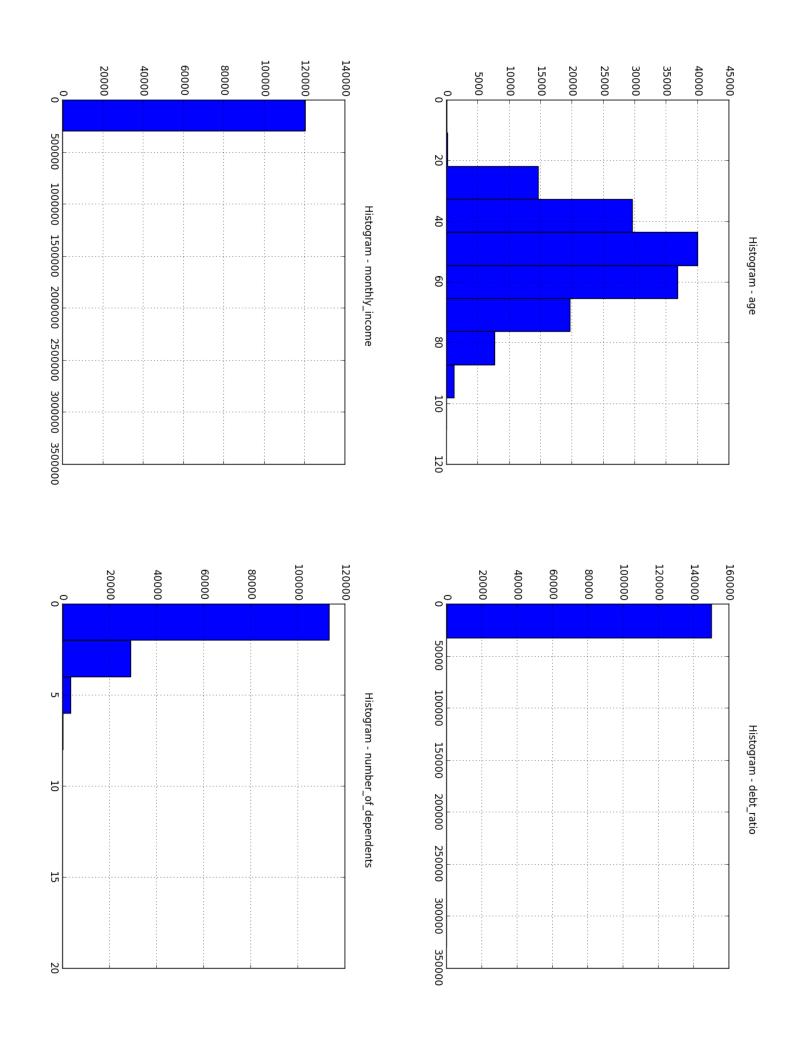
Recall score is: 0.04388459975619667 Precision score is: 0.5242718446601942

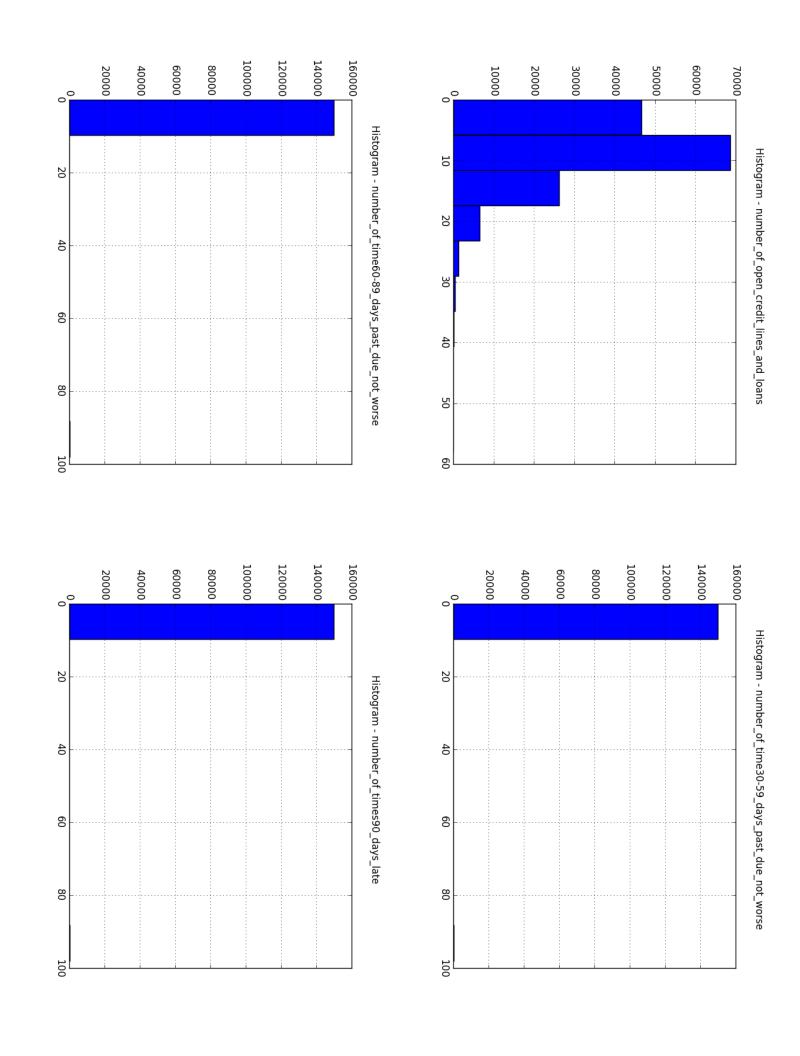
KNeighborsClassifier

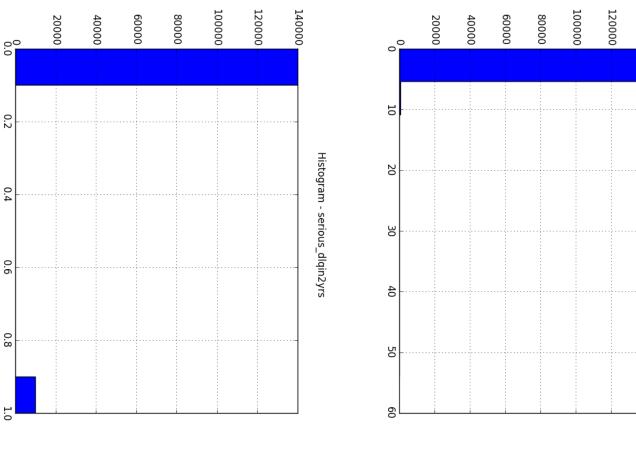
Accuracy score is: 0.9329066666666667 Recall score is: 0.025599349857781388 Precision score is: 0.34806629834254144

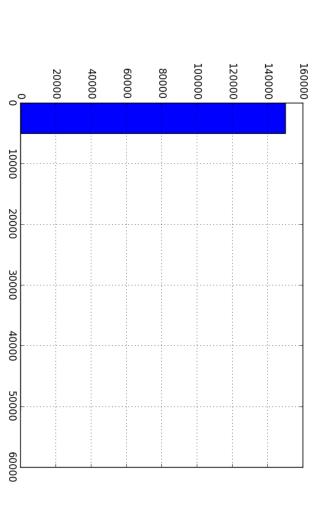
DecisionTreeClassifier Accuracy score is: 0.9

Recall score is: 0.28199918732222673 Precision score is: 0.25924542398206946









160000 _[

Histogram - number_real_estate_loans_or_lines

Histogram - revolving_utilization_of_unsecured_lines

140000