Credit Card Approval Prediction By using Classification Method

Zhengyuan Liu, Ruike Xu, Steven Ge

MATH 509

Professor: Jay Newby

Outline

1. Introduction of background and Dataset
2. Data cleaning
3. Feature selection
4. Modeling Assumptions
5. Modeling
    1. Logistic regression
    2. Decision Tree
    3. RandomForest
    4. Neural network
    5. SVM
6. Conclusion and Discussion

Introduction

- The credit score is an essential determination of personal credit evaluation.
- When proceeding with credit card applications, credit companies employ the applicant's personal information and credit history to evaluate future defaults and overdue payments.
- Our objective is to classify whether the applicant is a good credit holder that would repay on time

Data Acquisition

- The dataset we used contains transactions made by credit cards in September 2013 by European cardholders.
- Variables includes marriage status, property ownership, number of family member, number of children, annual income, etc.

Data Cleaning and Feature Selection

- Adjust credit status of applicants to binary categorical variable using scoring system
- Convert categorical predictors using one-hot encoding, adjust employment, births, and work age ratio

Data Cleaning and Feature Selection

- Conduct feature selection by using correlation
- Split data randomly into training 80% and testing 20%

Modeling Assumption

- Classification methods need to have discrete response variable
- Confounding effect
- Independent observations
- Logistic regression assumptions
  - Appropriate outcome structure
  - Linearity of independent variables and log odds
  - Lack of strongly influential outliers
  - Absence of multicollinearity
  - Assumption of a large sample size

Modeling

Logistic Regression:

Logistic regression is a type of generalized linear regression often used as a classification model to predict binary variables.

Formula: $$P(y=1) = \frac{1}{1+e^{-(\beta 0+\beta 1X1+\beta 2x2+\ldots\ldots+\beta nXn)}}$$

Confusion matrix (training):

Accuracy Score is 0.51703

	0	1
0	5389	4664
1	5047	5007

Confusion matrix (testing):

Accuracy Score is 0.51502

	0	1
0	1369	1144
1	1294	1220

Decision Tree:

Decision tree is a tree-like model, which is a decision-making process based on multiple decisions.

Random Forest:

Random forest is a supervised learning model based on a series of decision trees. The random forest combines different decision trees, and produces the outcome by voting the majority of all decision trees.

Confusion matrix (training):

```
Accuracy Score is 0.80067
      0     1
0  8380  1673
1  2335  7719
```

```
Accuracy Score is 0.80067
      0     1
0  7933  2120
1  1888  8166
```

Confusion matrix (testing):

```
Accuracy Score is 0.66382
      0     1
0  1778   735
1   955  1559
```

```
Accuracy Score is 0.66421
      0     1
0  1647   866
1   822  1692
```

Neural network:

Neural network is a computational learning system that uses a network of functions to understand and translate a data input of one form into the desired output. (Hidden layers 200, 150, 100, 50)

SVM:

Support vector machine algorithm is to find a hyperplane in an N-dimensional space that distinctly classifies the data points

Confusion matrix (training):

```
Accuracy Score is 0.64893
      0     1
0  6025  4028
1  3031  7023
```

```
Accuracy Score is 0.54339
      0     1
0  5502  4551
1  4630  5424
```

Confusion matrix (testing):

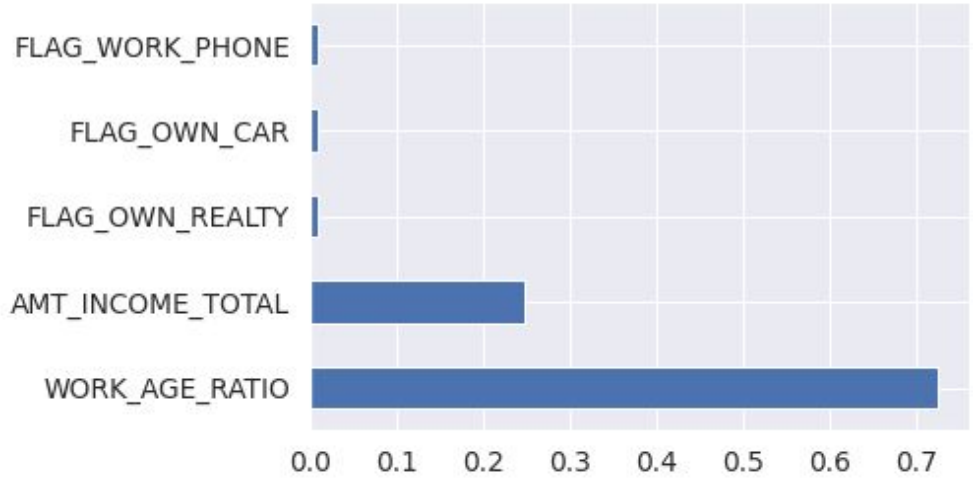
```
Accuracy Score is 0.56734
      0     1
0  1307  1206
1   969  1545
```

```
Accuracy Score is 0.52457
      0     1
0  1349  1164
1  1226  1288
```

Conclusion

- Random forest has the best performance
  - Accuracy Score: 0.66421

	precision	recall	f1-score	support
0	0.67	0.66	0.66	2513
1	0.66	0.67	0.67	2514
accuracy			0.66	5027
macro avg	0.66	0.66	0.66	5027
weighted avg	0.66	0.66	0.66	5027



Discussion

- Optimize data sources by using better sampling methods.
- Use some ways to avoid overfitting problem, like cross-validation.
- Find more relative data from other countries
- Using more recent data may improve the utility of our model on current society

- Thank You for Listening