PREDICTING CREDIT CARD DEFAULTERS

By Taha Shahid

ABSTRACT

Machine learning techniques can be used as an important tool for financial risk management.

Data Science Career Track - Springboard

1. Introduction

- The data set selected is called “Default of credit card clients Data Set” .
- Predict which classification method / data mining technique that will give the best accuracy for the probability of default of credit card clients.
- Client will be banks in Taiwan because banks will be utilizing the analysis and machine learning models to better understand their clientele and will be able to make decisions on their customers which will benefit them from customers that will most probably default.

2. Classifier set

- Decision Trees
- Random Forest
- Logistic Regression
- Ensemble Techniques Boosting
  - XGBoost
  - ADABoost
  - Voting Classifier
- Bagging

3. Data Description

- Age (in years);
- Gender (where 1 = male and 2 = female);
- Marital status (where 1 = married, 2 = single, and 3 = other);
- Education (where 1 = graduate school, 2 = university, 3= high school, and 4 =other); and
- Credit limit.

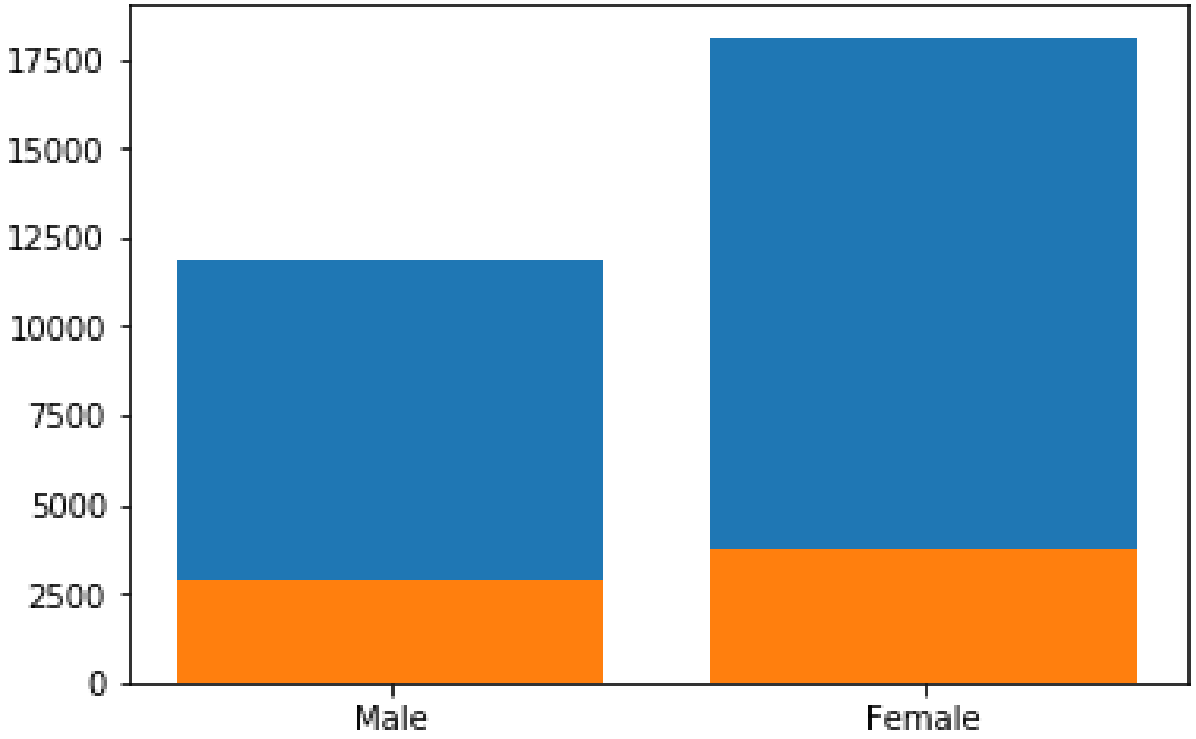
3. Data Description (CONT'D)

- Past payment status that month, where:
  - -2 = no consumption;
  - -1 = paid in full;
  - 0 = the use of revolving credit;
  - 1 = payment delay for one month;
  - 2 = payment delay for two months;
  - …;
  - 8 = payment delay for eight months;
  - 9 = payment delay for nine months and above;
- Amount billed that month; and
- Amount paid that month.

4. Statistical Analysis

Defaulters based on Gender

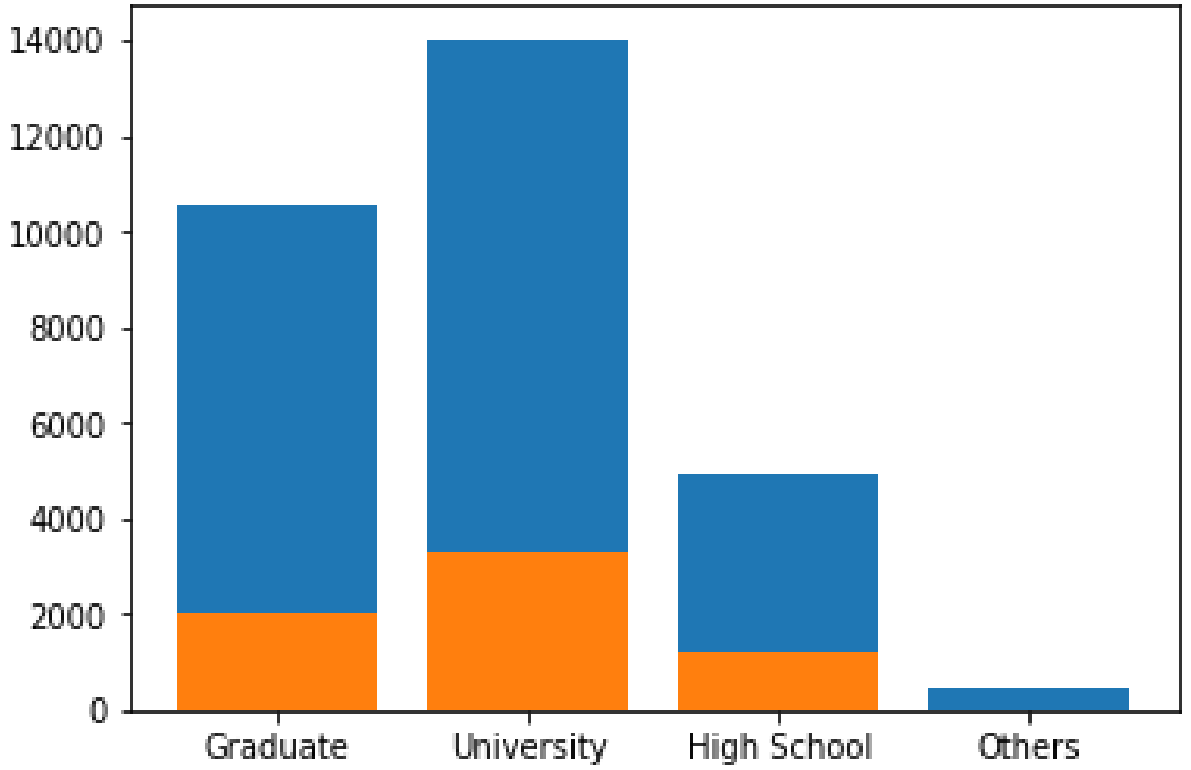
- Proportion of male defaulters is 24.16 %
- Proportion of female defaulters is 20.77%
- Total number of males in data was 11888 and total number of females in data was 18112.



4. Statistical Analysis (CONT'D)

Defaulters based on Education Level

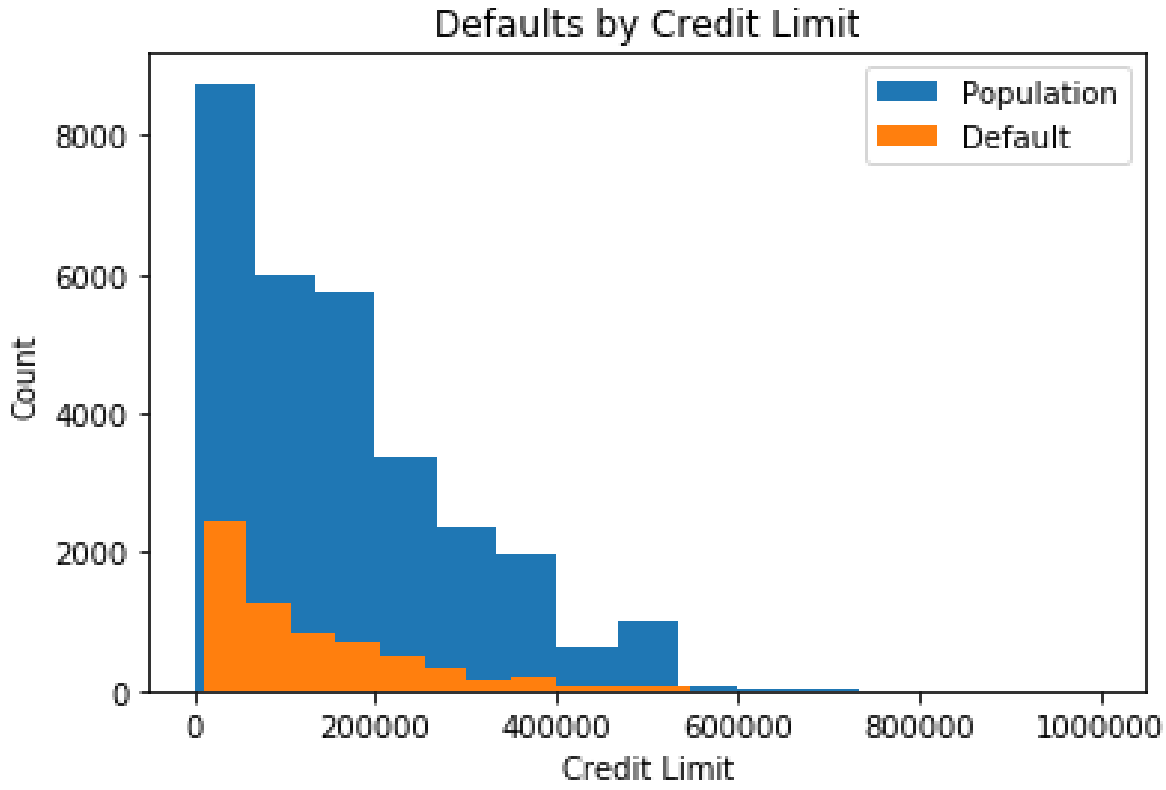
- University level with 23.74%
- High School level with 23.45%
- Graduate level with 19.24%
- Others with 5.7%



4. Statistical Analysis (CONT'D)

Defaulters based on Credit Limit

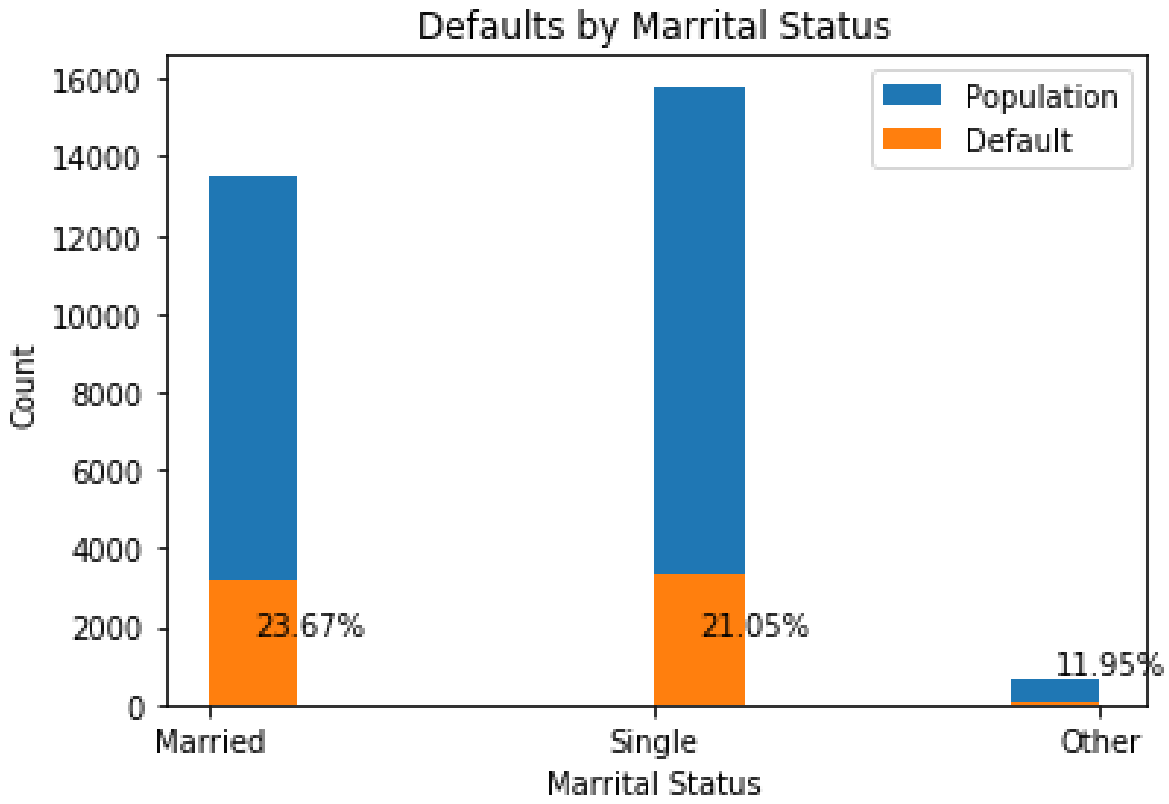
- From 0 limit balance to 200000: 25.34%
- from 200000 limit balance to 400000: 15.48%
- from 400000 limit balance to 600000: 12.28%
- from 600000 limit balance to 800000: 7.69%



4. Statistical Analysis (CONT'D)

Defaulters based on Gender

- Married with 23.67%
- Single with 21.05%
- Others with 11.95%



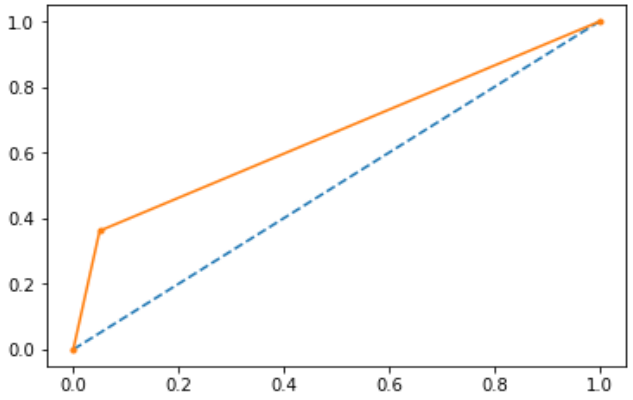
5. Feature Engineering

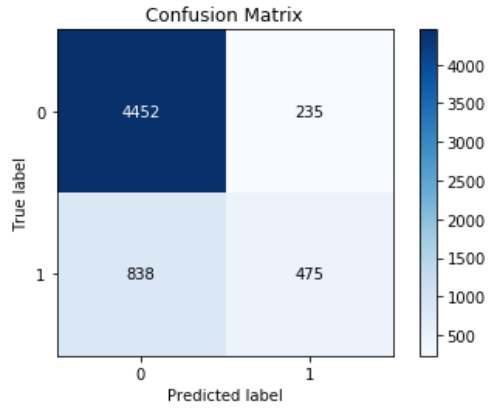
- Feature 1:
  - Addition of first three payments names PAY_0, PAY_2 and PAY_3.
- Feature 2:
  - Using above three payments and making a binary column if any of the three above payments had -2 or +2.
- Feature 3:
  - Sixth month payment minus the sixth month bill amount.
- Feature 4:
  - First month payment minus the first month bill amount.

5. Machine Learning Results

Decision Trees

- Decision tree structure is composed of nodes and leaves
- Root node is the top-most node in the tree
- Decision tree result in simple classification rules and handle nonlinear and interactive explanatory variables
- It scored ROC-AUC score of 0.6558
- Test Accuracy Score : 82.11%





5. Machine Learning Results

Random Forest

- An ensemble learning method for:
  - Classification
  - Regression
  - Miscellaneous
- Operates by:
  - constructing a multitude of decision trees
  - training time and
  - outputting the class that is the mode of the classes (classification) or mean prediction (regression)

- ‘GINI’ criterion was used with 4 jobs and 400 estimators

- Test Accuracy Score : 82.08%

5. Machine Learning Results
Logistic Regression

- Used in:
  - Credit Risk Modeling
  - Financial Prediction
- Operates by:
  - Studies association in categorical dependent variable and set of independent variables
  - Produces a probabilistic formula of classification
  - Deal with non-linear effects of explanatory variables
- Parameters:
  - penalty='12', dual=False, max_iter=145, C=2, fit_intercept=False
- A test set score was 78.11%.

5. Machine Learning Results

XGBoost and ADABoost

- Ensemble learning method to decrease variance in bias
- Boosting is a machine learning meta-algorithm
- Converts weak learners to strong ones
- A weak learner is defined to be a classifier that is only slightly correlated with the true classification

- Test accuracy score of XGBoost : 82.02%
- After feature engineering XGBoost : 82%

- Test accuracy score of ADABoost : 81.93%

5. Machine Learning Results
Ensemble and Bagging

- An ensemble learning method for:
  - decrease variance (bagging)
  - bias (boosting)
  - improve predictions (stacking)
- Bootstrap Aggregation(Bagging):
  - reduce the variance for those algorithms that have high variance
  - classification and regression trees (CART)
  - Bagging is the application of the Bootstrap procedure to a high-variance machine learning algorithm, typically decision trees

- Ensemble's voting classifier chose XGBoost with default parameters as the best Classifier with Test Accuracy of 81.8%
- Bagging with Random Forest achieves test accuracy of 81.95%
- Bagging with Decision Tree achieves test accuracy of 81.53%

5. Machine Learning Results

All Results

Models/Classifiers	Test Accuracy (%)
Decision Tree	82.11
Random Forest	82.08
Logistic Regression	78.11
XG-Boost	82.00
ADA-Boost	81.93
Ensemble (Voting Classifier)	81.8
Bagging	81.53

Thank You!

By Taha Shahid