Predicting Default of Credit Card Clients

By Tsolmon Jargalsaikhan

Data Set Information

UCI data set contains23 predictive variables and 30K instances. The research employed a binary variable, default payment (Yes = 1, No = 0), as the response variable, and used the following 23 variables as explanatory variables:

X1: Amount of the given credit (NT dollar)

X2: Gender (1 = male; 2 = female).

X3: Education (1 = graduate school; 2 = university; 3 = high school; 4 = others).

X4: Marital status (1 = married; 2 = single; 3 = others).

X5: Age (year).

Data Set Information

X6 - X11: History of past payment. We tracked the past monthly payment records (from April to September, 2005) as follows: X6 = the repayment status in September, 2005; X7 = the repayment status in August, 2005; . . .;X11 = the repayment status in April, 2005. The measurement scale for the repayment status is: -1 = pay duly; 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.

X12-X17: Amount of bill statement (NT dollar). X12 = amount of bill statement in September, 2005; X13 = amount of bill statement in August, 2005; . . .; X17 = amount of bill statement in April, 2005.

X18-X23: Amount of previous payment (NT dollar). X18 = amount paid in September, 2005; X19 = amount paid in August, 2005; . . .;X23 = amount paid in April, 2005.

Client and Business Problem

Our client, Bank in Taiwan, would like to predict which customers will default next month based on the data set provided.

By predicting potential defaults early the bank can take measures to limit the exposure and mitigate potential losses.

Data Wrangling Steps

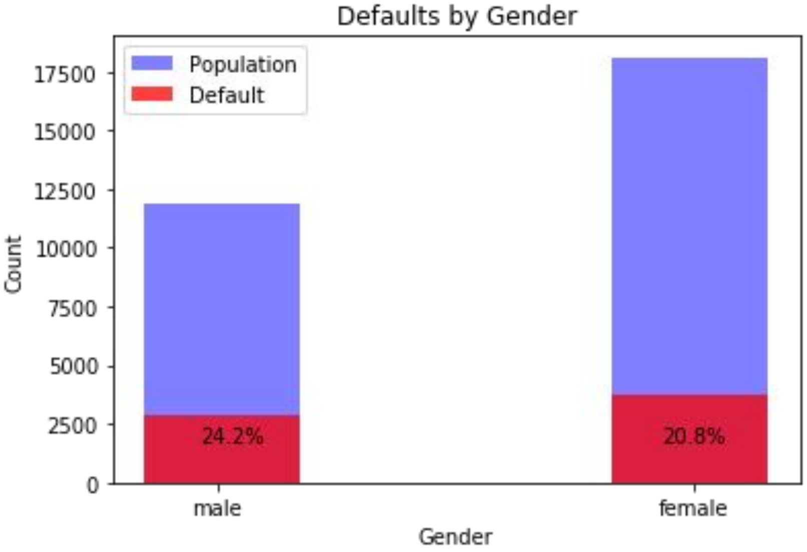
In general, this data set was pretty clean and it didn't require major cleaning efforts.

- Dropped one of the header rows
- Education column had one extra categories 0, 5, 6 which had only 345 values so I combined with category 4 which is an other category.
- Marriage column had undefined 0 category which had only 54 rows so I merged it to 3 which is an other category.
- Changed data types of columns with categorical values (gender, education, marriage status).

Exploratory Data Analysis

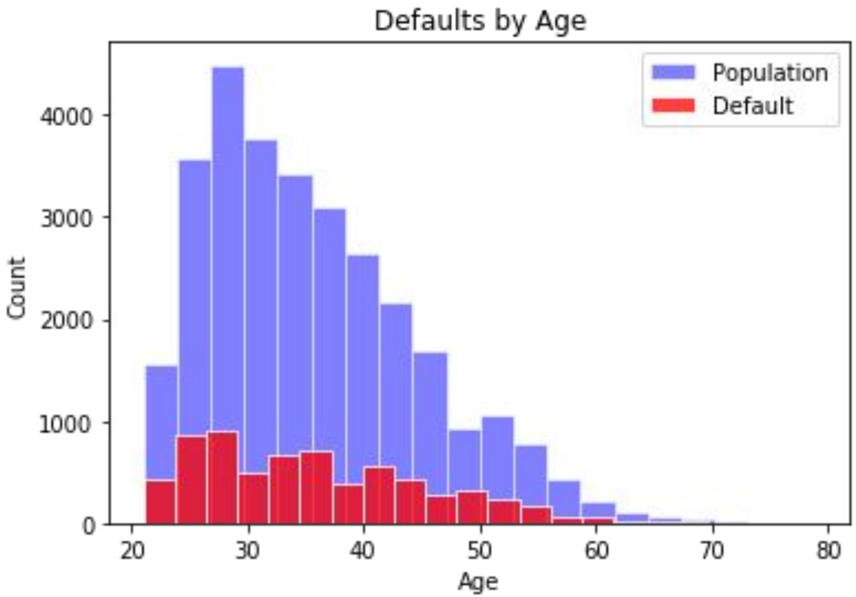
Notable findings:

- As of the latest month, customers that are on default status (1 or more months behind payment) were approx. 22.7%.
- Male customers have higher default rate of 24.16% compared to female customers' default rate of 20.78%.



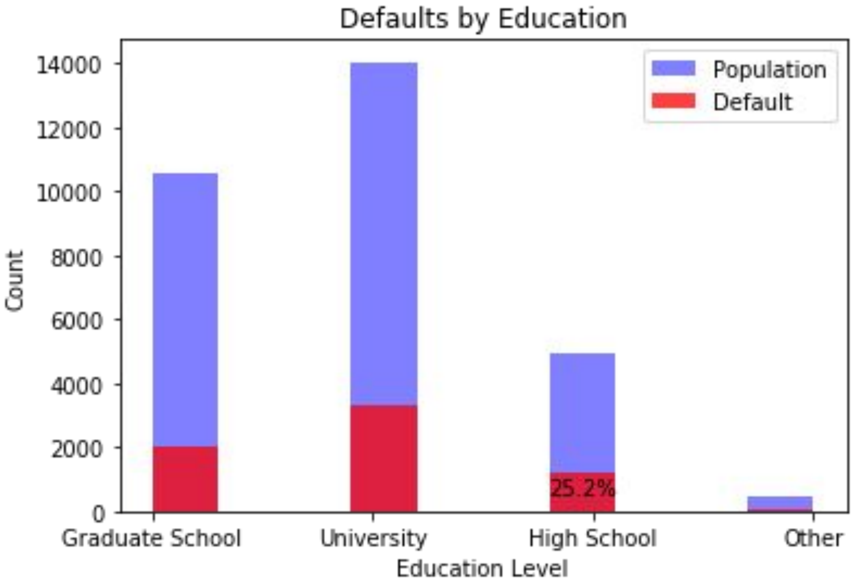
Exploratory Data Analysis

- Customers under age of 25 have a higher default rate of 27.19% than customers aged over 25 which has a default rate of 21.45%.



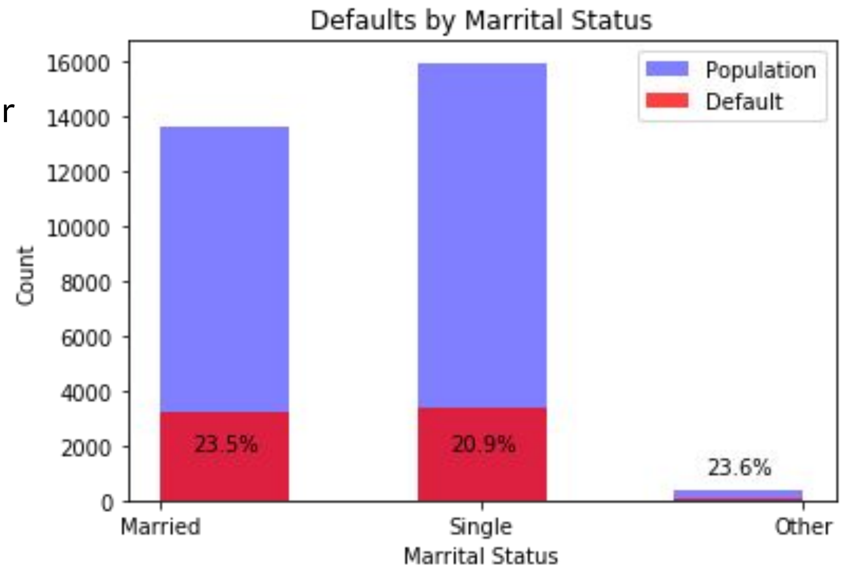
Exploratory Data Analysis

- Customers with high school degree have highest default rate of 25.16% than other education categories.



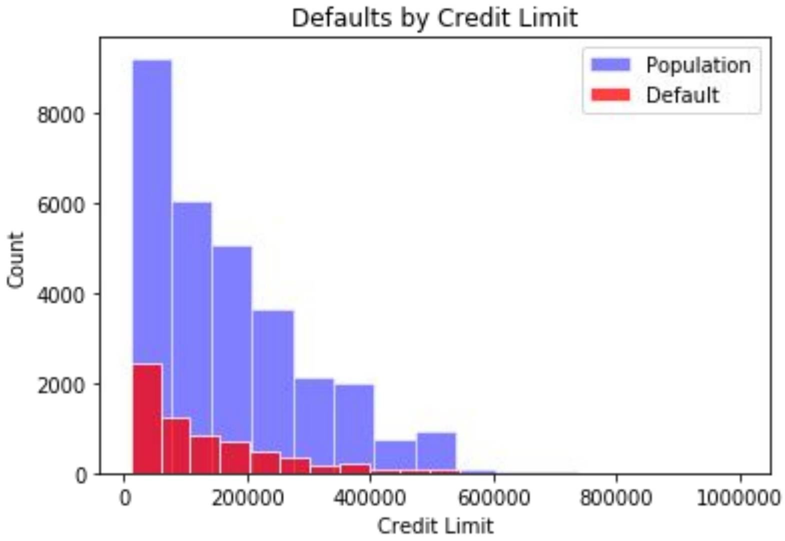
Exploratory Data Analysis

- Married and other category customer had higher default rate than single customers.



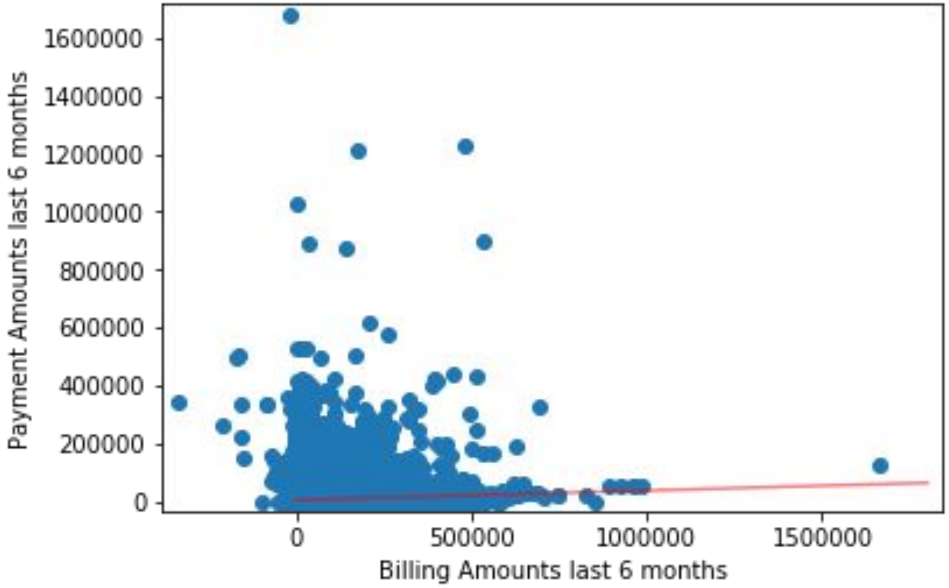
Exploratory Data Analysis

- Defaults by Credit Limit is not normally distributed.



Exploratory Data Analysis

- Billing Amounts and Payments Amounts are positively correlated; there is gradual increase of Payment Amounts as Billing Amounts increase.



Predictive Model Building

Performance of classifiers used on the data set:

- Tuned Logistic Regression Accuracy: 80%
- KNN Classifier Accuracy: 78%
- SVM Accuracy: 77%
- Decision Tree Classifier Accuracy: 72%
- Random Forest Accuracy: 82%
    - After using SMOTE my recall rate improved by 4% which is good in this case because predicting defaults is much more important.

Predictive Model Building

- **Deep Learning with Keras:**

```
Train on 21000 samples, validate on 9000 samples
Epoch 1/20
21000/21000 [==============================] - 12s 554us/step - loss: 3.6849 - acc: 0.7714 -
val_loss: 3.2953 - val_acc: 0.7956
Epoch 2/20
21000/21000 [==============================] - 5s 222us/step - loss: 3.6811 - acc: 0.7716 -
val_loss: 3.2953 - val_acc: 0.7956
Epoch 3/20
21000/21000 [==============================] - 5s 224us/step - loss: 3.6811 - acc: 0.7716 -
val_loss: 3.2953 - val_acc: 0.7956
```

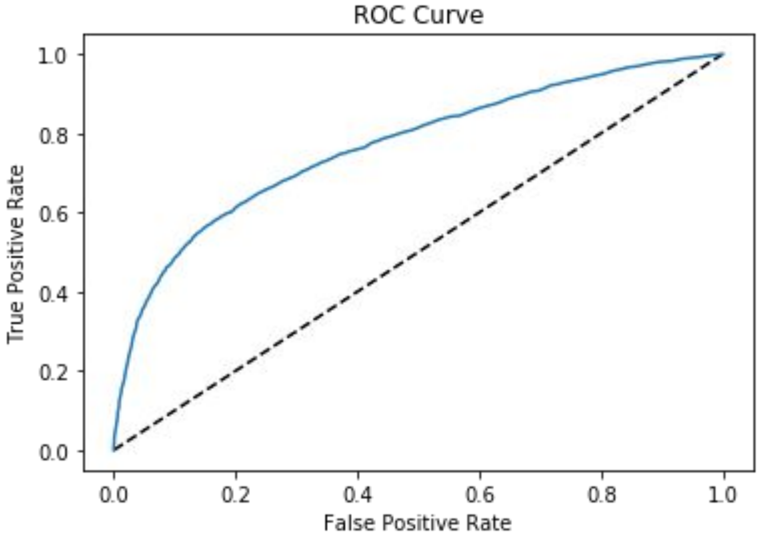
Conclusion

In conclusion, using Random Forest Classifier with SMOTE yields the best result with Accuracy Score of 82% and recall rate of 42%. Improvement in recall is important in this case because our customer will most likely care more about predicting actual defaults (TP).

```
             precision    recall  f1-score   support

          0       0.85      0.93      0.89      9337
          1       0.63      0.42      0.50      2663

avg / total       0.80      0.82      0.80     12000
```


ROC Curve
True Positive Rate
False Positive Rate