Guido Rossum, Senior Data Scientist

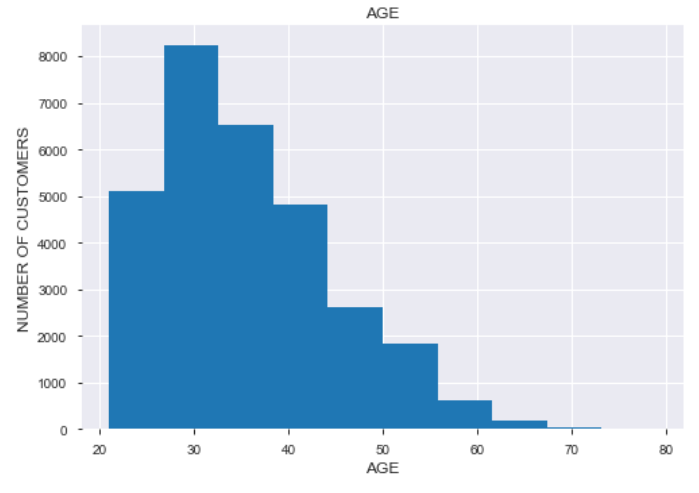
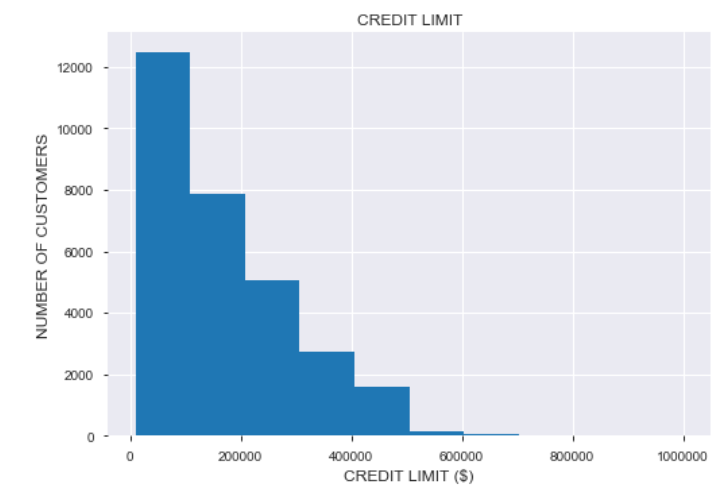
Data Science Team

Over the past year, Credit One, a credit scoring service company, has seen an increase in the number of customers in Taiwan who have defaulted on loans they have secured from various partners. As a result, Credit One, who is responsible for approving customers for loans, is at risk of losing business if the problem is not solved in a timely fashion. Our data science team was provided with the following information to analyze and build a predictive model to better assign customer credit limits:

1. 30,000 customer credit card accounts
2. Amount of current credit limit (individual + family, if any) ranging from $10k to $1million
3. Demographic information: gender, marital status, age, education level
4. 6 months of credit status information (see legend table)
5. 6 months of billing statements and payment amounts from Apr-Sept 2005
6. “default next month” status (default, not default)

The following report summarizes the data science process implemented, visualizations, observations, and recommendations going forward. As the dependent variable, “credit limit”, is continuous, regression is chosen as the modeling algorithm. Initial exploratory data analysis within Jupyter notebook consisted of data upload, checking for data types (integers/objects), renaming some columns, removing unnecessary columns, removing duplicate entries, and checking for missing values. It is determined that all data types are integers, except for the object columns of “sex”, “education”, and “default”. “Sex” and “default” have only 2 unique entries and are converted to binary values using LabelEncoder. Education, with 4 unique entries, is converted to separate binary columns using OneHotEncoder. A legend table is shown below to describe the monthly credit status data for each customer. The negative values of -2 and -1 are indicative of customers that are paying on-time. To ensure that customers who are paying on-time are not penalized within the modeling algorithm, all values of -2 and -1 are converted to zero values.





The histogram of credit limit for all customers reveals that most customers have credit limits in the lower range and almost entirely below $500k. Customer ages are mostly below 55 years.

A correlation matrix determines that the following variables are not relevant in this problem and are removed: pay status, sex, marital status, and education (sans graduate school). Three regression models are selected and result in the following accuracies:

1. Random Forest Regressor: 45%
2. Linear Regression: 24%
3. Support Vector Regression: 33%

Random Forest Regressor is selected as the algorithm of choice. However, feature selection options and tuning parameters are unsuccessful in improving model accuracy above 50%. It was decided that the dependent variable, “credit limit”, will be separated into bins so that a Classification Algorithm can be performed.

Initially, credit limit is separated into 5 bins as follows: $0-$200k, $200-$400k… $800k-$1million. Random Forest Classifier, Gradient Boosting Classifier, and Support Vector Classifier all return model accuracies of ~73%. Random Forest Classifier was chosen as the model of choice. Additionally, a feature importance algorithm indicated that age, September bill, and August bill were the most important features contributing to the modeling results.

An additional trial was performed with credit limit separated into 10 bins representing $100k increments up to $1 million. However, this increased differentiation between customers into 10 groups resulted in a Random Forest Classifier model accuracy of 57%, significantly below a desired outcome.

The issue that is presented within the 5-bin classification scenario is that most customers are represented between only 2 bins ($0 -$200k and $200k - $400k). This will not likely provide clients with enough customer differentiation in assigning new credit limits. If all customers within a $200k credit limit margin are treated equally, how will their individual credit limits be determined? It appears that in order to produce a better predictive model, additional customer information is needed such as: salary, credit score, employment history, and outstanding debt, among other relevant information.