**Credit One - Data Science**

Over the past year, Credit One has experienced an increasing number of customers defaulting on loans. Credit One, as the credit scoring service approving the loans, risks losing business if the problem is not solved.

Credit One has requested that we develop an empirically sound data model to answer two fundamental questions:

* How do you ensure that customers can/will pay their loans?
* Can we approve customers with high certainty?

**Overview of data**

Provided with 3670 customer records with 24 features

* Demographic data
  + Customer Sex
  + Education
  + Marital Status
  + Age
* Limit balance
* Defaulted status
* 6 months of payment information
  + Payment actions
  + Billed amount
  + Paid amount

**Exploratory Data Analysis**

* 78.3% of current clients will not be defaulting payment next month
* 21.7% of clients will be defaulting next month
  + This metric is baseline to reduce for Credit One
* From the continuous variables of age and limit balance, limit balance had the strongest correlation to default rates
* From the following variables, the correlation order from strongest to weakest is: limit balance > education > age > sex > marital status
* Out of the variables for repayment status, bill amount, and previous payment, repayment status had the strongest relationship to defaulting rates

**Regression Modeling**

Regression and classification modeling was used to develop a model suitable for predicting how much credit to approve for a customer.

Three regression models were used to predict default status and limit balance. None of these models could be used as suitable predictors. Default status is binary, so could not be predicted accurately using a regression model. Using all features as independent variables to predict limit balance yielded the best predictors using regression modeling, with the following predictors

| **Model** | **Score** |
| --- | --- |
| Random Forest Regression | 0.434 |
| Linear Regression | 0.361 |
| Support Vector Regression | -0.051 |

Regression modeling is not suitable for either default status or limit balance prediction. We then pivot towards classification modeling for predictions.

**Classification Modeling - Limit Balance**

Three classification models were evaluated as a predictor of credit limit balance for a customer.

| **Model** | **Score** |
| --- | --- |
| Decision Tree Classifier | 0.367 |
| Random Forest Classifier | 0.479 |
| Gradient Boosting Classifier | 0.469 |

None of these models were good predictors for limit balance, thus, we cannot accurately predict credit limit balance to approve for a customer application.

**Classification Modeling - Default Status**

Three classification models were evaluated as a predictor for default status of a customer.

| **Model** | **Score** |
| --- | --- |
| Decision Tree Classifier | 0.715 |
| Random Forest Classifier | 0.800 |
| Gradient Boosting Classifier | 0.808 |

Using all features as independent variables with default status as the dependent variable, we were able to achieve the best prediction modeling scores for classification. Gradient boosting classifier was the best model and yielded the following:

|  | Precision | True Positive Rate | False Positive Rate |
| --- | --- | --- | --- |
| Default | 64% | 31% | 69% |
| Not Default | 84% | 96% | 4% |

From this model we are able to accurately identify if an applying customer will not default 84% of the time and if a customer will default 64% of the time. The true positive rates for not defaulting applying customers is high at 96% ensuring a higher ROI for Credit One by approving these applying customers.

If only demographic data is available for evaluating a customer during the application process, the following scores were obtained using only demographic data as independent variables with default status as the dependent variable which are good (but not as good as using all features).

| **Model** | **Score** |
| --- | --- |
| Decision Tree Classifier | 0.745 |
| Random Forest Classifier | 0.743 |
| Gradient Boosting Classifier | 0.780 |

**Recommendation**

Credit One should utilize the Gradient Boosting Classifier model using all features available from the applying customer to determine if they should be approved for a loan based off of their defaulting status prediction in that model. If payment status, bill amount, and paid amount information is not available from the applying customer, utilizing demographic data would allow for correct prediction of non-defaulting customers 80% of the time.

**Questions to Investigate**

* How do you ensure that customers can/will pay their loans?

We are unable to ensure customers can/will pay their loans from data. We are only able to best predict if a customer is likely to pay off credit balances using accurate data available. Even with the best data and models, there are still false positives that were identified as customers who won’t default that may default in the future based on their personal circumstances. We are able to reduce the number of defaulting customers by more accurately identifying customers who will not default at time of their application.

* Can we approve customers with high certainty?

We are able to predict 84% of the time whether an applying customer will not default using all features as data for the model. This prediction score is reduced to 80% if only demographic data is available from the applying customer.