# Credit One – Customer Default Identification Report

**Project Summary:**

The scope of this project was to evaluate customer default rates and determine possible solutions using machine learning to predict which customers are most likely to default on their loans. Based on the Data Science framework that was provided at the start of the project, we applied a rigorous methodology to clean and pre-process the data and perform exploratory data analysis. Some formatting changes were necessary (changing data types, one-hot encoding, etc.) to prepare the data for modeling. Based on the EDA, we determined which machine learning models would be most appropriate for our use case. We created four models and performed cross-validation and confusion scoring to evaluate model effectiveness.

**Exploratory Data Analysis – Findings:**

Initial evaluation of the data was unrevealing, and it was difficult to determine if any one feature of the data had predictive value. However, a PhiK correlation analysis of the data showed an interesting interaction between customer payment and billing history. (see below)

**Timeline

Description automatically generated with medium confidence**

Additionally, we noted a strong correlation between balance limit and billing history and a moderate correlation between balance limit and use of revolving credit (PAY\_0 – PAY\_6). There is also a moderate correlation present between customer default and use of revolving credit.

Analysis of customer demographic data did not show any one customer feature that was telling of default behavior. In looking at default rates by customer age for example, there was a similar pattern between the distribution of the customer’s ages and rates of default. (see below)

Chart, bar chart, histogram

Description automatically generated

Chart, line chart

Description automatically generated

This pattern remained similar when evaluating customer counts and default rates across other demographic features such as marriage status, education level and gender. (see example below)

Chart, bar chart

Description automatically generated

Chart, bar chart

Description automatically generated

**Machine Learning Modeling:**

We evaluated four machine learning algorithms using our data. (Random Forest Classifier, Gradient Boosting Classifier, Decision Tree Classifier and Support Vector Classifier) All four models performed similarly, with the Random Forest and Gradient Boosting classifiers performing equally well. Validating the models with the test data showed a weighted average accuracy score of 0.80 and 0.81 respectively. Confusion matrix pattern was also similar between the two models.

**Conclusions:**

While we were provided a robust dataset, there were no obvious features in the customer data that provided us with a definitive method for determining credit worthiness. Rather, evaluating parameters such as payment history and the use of revolving credit (especially the initial payment month based on correlation analysis) seem to be the most telling. There is only a weak correlation between credit limit and default rate, so that relationship in of itself is not predictive. Some additional parameters in the customer data such as income level and prior history of loan default might be more revealing. Creation of a “risk index” based on the available data might help establish a more linear relationship with respect to default behaviors. Finally, bringing in some additional data points, along with leveraging feature engineering techniques may also prove insightful.

Our machine learning models showed good predictive value based on validation scoring, and either the Random Forest Classifier or Gradient Boosting Classifier could be used to implement a solution to predict risk of customer default. I would encourage CreditOne to pursue data engineering initiatives along these lines.