Customer Default Identification Report

Credit One

Thomas Higginbotham

# Problem:

Credit One has experienced an increase in default rates, which is bad for Credit One and is likely to result in the loss of Credit One’s business customers.

# Investigation:

Guido Rossum from Credit One has provided some data showing customer payment history for the past 6 months, including demographic data (age/gender/education/marital status), the next bill amount, the next payment amount, the payment status, the credit limit, and whether the payment next month is default or not.

Guido Rossum has asked that I use this data to investigate the following:

1. How can it be ensured that customers can/will pay their loans?
2. Can customers be approved with high certainty?

# Data Science Process Framework:

The Zumel and Mount data science framework was used in the examination of the data and investigation of the questions posed by Guido Rossum. This process was chosen because both the goals and the data collection/management process were both well-defined. Additionally, we have prior experience building, evaluating, critiquing, and presenting decision models.

## Define the Goal

The goals of this investigation are to determine if models can be built that will ensure that customers can pay their loans, and approve customers with high certainty.

## Collect and Manage Data

The data was provided to us via a mySQL database maintained by Credit One. Specifically, we were provided access to the credit table, containing 3668 observations for the the following:

|  |
| --- |
| LIMIT\_BAL |
| SEX |
| EDUCATION |
| MARRIAGE |
| AGE |
| PAY\_0 |
| PAY\_2 |
| PAY\_3 |
| PAY\_4 |
| PAY\_5 |
| PAY\_6 |
| BILL\_AMT1 |
| BILL\_AMT2 |
| BILL\_AMT3 |
| BILL\_AMT4 |
| BILL\_AMT5 |
| BILL\_AMT6 |
| PAY\_AMT1 |
| PAY\_AMT2 |
| PAY\_AMT3 |
| PAY\_AMT4 |
| PAY\_AMT5 |
| PAY\_AMT6 |

Relatively little cleaning was needed for this data. PAY\_0 was renamed to PAY\_1, and there was a duplicate header line deep within the data which was also removed. Duplicate records were dropped. Additionally, all categorical values were converted to integers (for example, “male” and “female” became 0 and 1), and all values were converted from object data types to integer data types for further analysis. Following the data import and cleansing, 2396 observations remained.

## Build the Model

Four models total were built. The first two attempted to address the two questions provided by Guido Rossum using a regression model, and the last two attempted to address those questions using a classification model.

Feature selection was validated using a SelectKBest selector, comparing the Univariate score (-*Log*(*Pvalue*)), the SVM weights, and SVM weights after selection.

Chart, histogram

Description automatically generated

This methodology allowed me to choose the features with the highest weighted p-values and ignore features that were not relevant to solving the questions.

## Evaluate/Critique the Model

At first, the questions were evaluated using various regression models (Random Forest Regressor, Linear Regression, and SVR. The performance of each regression model was then evaluated, producing the following r2 scores:

### Predict Bill Payment (Regression)

Random Forest Regressor 0.09970660000455785

Linear Regression 0.07242378022493834

Support Vector Regression -0.08149319567545306

With the input data set, none of these models will perform well to ensure that customers will be able to pay their loans. Of the 3 options, Random Forest Regressor should perform the best as it has the highest r2 value.

### Predict Credit Approval (Regression)

Random Forest Regressor 0.4455740660586594

Linear Regression 0.38293506444885894

Support Vector Regression -0.03954910243741098

With the input data set, none of these models will perform well to ensure that credit limits are predicted. Of the 3 options, Random Forest Regressor should perform the best since it has the highest r2 value.

### Predict Bill Payment (Decision Tree Classification)

Overall Classification

precision recall f1-score support

0 0.21 0.22 0.21 93

1 0.81 0.81 0.81 387

accuracy 0.69 480

macro avg 0.51 0.51 0.51 480

weighted avg 0.69 0.69 0.69 480

Overall Classification Score: 0.9386477462437396

Made into a classification problem and using a decision tree classifier, the precision, recall, and f1-scores are all .8 or above, and the overall r2 score is .93. This model should be able to predict whether customers can make their payments or not.

### Predict Credit Approval (Decision Tree Classification)

Overall Classification

precision recall f1-score support

0-50000 0.44 0.40 0.42 57

140000-200000 0.35 0.38 0.36 122

200000-1000000 0.50 0.49 0.49 127

50000-140000 0.55 0.54 0.54 174

accuracy 0.47 480

macro avg 0.46 0.45 0.46 480

weighted avg 0.47 0.47 0.47 480

Overall Classification Score: 0.8927378964941569

The overall model score (r2) of .89 indicates that this model should be good at predicting the credit limit. However, the lower precision, recall, and f1 scores indicate that the model might be useful only in some situations.

## Present Results and Document

Overall, it appears that a classification model is best to solve the questions posed by Guido Rossum. Regression models using the data provided do not seem to be able to achieve the accuracy required. It should be easier to predict whether customers can pay their loans using the classification models. For predictions about whether a customer should be approved or not, using classification models for specific amounts for different demographic and payment features may be the best choice.