Credit One

Credit Approval Service

Over the past year, Credit One corporate clients have 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:

* Should a customer credit application be approved?
* How much credit should a successful applicant be provided?

# Chart, bar chart Description automatically generatedOverview of data provided

We were provided with 30,000 customer records with 24 features.

Customer demographics:

* Sex

Education

* Marital status
* Age
* Amount of credit extended
* Whether the customer had defaulted

6 months of payment information was provided on:

* Payment actions
* Amount billed
* Amount paid

Relevant information on the 30,000 records

* + Credit limit
    - Minimum - $10,000
    - Maximum - $1,000,000
    - Average - $167,484
  + Average customer age – 35.5 years
  + Average balance - $39,000 - $47,000
  + Average payment - $5,300

A thorough exploratory data analysis was performed and did not reveal any surprises in the data. There was no correlation in credit limit or default status with any of the demographic information.

# Predicting amount of credit

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

# Linear regression model

Linear regression was performed to predict how much credit to approve for a customer. Table 1 shows the criteria generally used to assess how well a linear regression model performs.

|  |  |
| --- | --- |
| Value as a predictor | Criteria |
| Good Predictor | Greater than 0.7 |
| Moderate Predictor | Between 0.5 and 0.7 |
| Poor Predictor | Less than 0.5 |

Table : Predictor Criteria

Three different regression models were employed and none of them provided results that would perform better than the model currently being used. As shown in Table 2, at best the model were moderate predictors.

|  |  |
| --- | --- |
| Model | Score |
| Random Forest Regressor | 0.467 |
| Linear Regression | 0.35 |
| Support Vector Regression | -0.05 |

Table : Model Performance

# Classification model – Amount of credit

Several classification models were evaluated for use in predicting how much credit to approve for a customer.

The two best performing models using this set of information performed at essentially 50%. This is no better than tossing a coin in determining how much credit to extend.

|  |  |
| --- | --- |
| Model | Score |
| Random Forest Classifier | 51.3% |
| Decision Tree Classifier | 40.6% |
| Gradient Boosting Classifier | 51.8% |
| Ada Boost Classifier | 47.9% |
| K Nearest Neighbors | 44.4% |

Bottom line: We cannot provide a model to effectively project how much credit to offer a customer.

# Classification model – Loan approval

We then used classification modeling to determine whether a customer should be approved for a loan, and the results were much better. Ran the model using all features provided and various subsets. Using the payment history features (PAY\_0 – PAY\_6) provided essentially identical results, accurately predicting whether a customer would default or not 82% of the time. This is a good result and can be used for making business decisions on whether to approve a loan.

|  |  |  |  |
| --- | --- | --- | --- |
|  | Correct Identification | False positive rate | Number of records |
| Default | 68% | 65% | 1989 |
| Not Default | 84% | 5% | 7011 |

# Recommendation

Credit One should use a Random Forest Classification model using past payment history as the basis of all features to determine whether to approve a loan application. The other features in the data set are not as useful in predicting whether someone will default on the loan. The most important feature is payment history.

Other models were evaluated, but showed signs of overfitting, and Random Forest provided the most consistent, reliable results.

# Questions to Investigate

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

This is not a data issue. At best, with the right set of data, we can predict who is likely to pay their loans, but we are unable to ensure a customer can and will pay their loans. Life is fluid and it changes. Someone who is a good credit risk can lose their job and become unable to repay. Not amount of analytics can ensure a customer will repay their loan.

1. Can we approve customers with high certainty?

Given the current set of data, Credit One can predict with greater than 80% accuracy whether a customer credit request should be approved. The key is to get past payment performance history on the customer.

# Lessons learned

Trust the data – During EDA, took too deep of a dive into the data itself, trying to understand each data field. In the end, when the data is cleaned and ready, accept the data as it is.

Data modeling – Some sets of data and dependent variables do not do well with different types of modeling. We were unable to use the data provided to predict how much credit to approve. However, we can predict with 82% accuracy whether someone will default on a loan.

Features – features that intellectually should be important in modeling may or may not be important. Need to perform enough analysis to assure understanding but be prepared to accept the answer the data provides.