Credit One Default Rates Report

Credit One is facing an increase in the number of customers with a defaulted loan in the past year. A data science team with data analytics techniques were employed to investigate factors (characteristics of customer data) affecting default rates. Using Credit One’s historical data consisting of 30,000 entries, the data science team will build an accurate model to help predict the next steps for Credit One regarding customers and their defaulting loans. This hopes to avoid a decrease in clients and maintain profits from them.

For building our models in the first phase, we used many types to configure the test and give us the best scoring: Linear Regression, Random Forest Regressor, Support Vector Regression. In the second phase of processing our data and selecting the features that will be tested in a different set of models, we used Nearest Neighbors Classification (KNeighbors), Gaussian Naive Bayes (GaussianNB), Decision Tree Classifier, Support Vector Model (SVC), and Logistic Regression. These particular models applied a cross-validation of 10 stratified k-folds (an improved version of k-1 folds for classification with imbalanced class distribution purposes). Based on the selected features and the scoring type, the second phase models had improved in performance and provided good evaluation scores. The second phase also included modifiers such as improving the models with ensembles such as Random Forest Classifier, Gradient Boosting Classifier, Extra Tree Classifier, AdaBoost Classifier, Voting Classifier and Bagging meta-estimator. Although the evaluation scores did improve, there was not a big difference significant enough.

To conclude which model will prove significance in the second phase, we tuned the models in order to provide predictions in our most superior model. Part of this process is to implement GridSearchCV which is tuning the hyper-parameters of an estimator and is optimized for cross validation scoring. The following models were tuned with GridSearchCV: Gradient Boosting, Logistic Regression, Decision Trees (with Randomized Search CV type), K-Nearest Neighbors, Support Vector Machine, and Random Forest from Phase 1 modeling. These models applied were then evaluated using accuracy scoring of the performance of cross-validation. Two out-performance models in phase 2 are listed below:

|  |  |  |
| --- | --- | --- |
| Algorithm | Gradient Booster | Support Vector Model |
| Accuracy Score | 0.8216 | 0.8222666666666667 |

Based on the trained models computed, Support Vector Modeling and Gradient Boosting Classifiers proved to be useful in this case given our target data. Classifiers in phase 2 modeling versus Regressors in phase 1 proved more useful given the imbalanced class distributions in the dataset and many discrete, categorical variables. As more data is provided and optimized about the targeted customer, it can be possible to prove the phase 1 models to provide more insight into each feature whilst increasing our accuracy in all our models.