Overview Readme

The following project was designed to explore financial modeling using machine learning techniques. The database employed was used to predict whether a

borrower would fully payoff the loan taken out. Each loan had the borrowers credit score, years of employment, credit history, annual income, purpose for

loan, months delinquent of payment and so on. The predictions were made by taking one attribute of the data base, known as the target and using the rest

(the features)of the columns to find patterns of which would reveal information. Loan Status, which had the value Fully Paid Off or Charged Off seemed like

the obvious choice for the target value. The next step of the processes was to organize the features into patterns of prediction where we could see the

whether the loan was paid off or charged off. Then test if our patterns of features were correct with new data that was not previously used(New target

values with their associated feature values). The most common way of modeling is to use a linear regression where a plot of a straight line best connects or

is in the area of the most data points. However when dealing with data that is binary or non-continuous the accuracy of linear regression becomes

inconsequential for making meaningful predictions. A logarithmic regression model would fit the criteria predicting the target value yet was only marked with

an improvement yet its accuracy is still in the sixties. A very basic yet powerful tool in machine learning is the decision tree, which allows a model to break

up data into variety of different groups down to the individual level. One of the most successful models constructed had eight-five percent accuracy was

utilizing a basic decision tree. There are numerous ways of making such decision tree classifiers as there are levels of sophistication of such models. The

most popular decision tree classifier used for regression analysis is currently Random Forest which gained its notoriety from being able to specialize each

tree for improved accuracy. However the Random Forest Model was only able to obtain an accuracy in the seventy-sixth percentile. The next general type of

model used is the Gradient Boosting Classifier which essentially attempts to make a bad prediction and learn from its mistakes. The more mistakes the

model makes and the slower it learns the better the levels of accuracy. XGBoost is big name in gradient boosting for those of us that haven't been reading

machine learning on Friday night. However with the dataset on hand while the XGBoost model scored two percent higher on the precision or the percentage

of true positives to the sum of true positives and true negatives, the recall of the XGBoost was more than twenty-five percent lower than the basic gradient

model (Recall is the true positives to the summation of true positives and false negatives). The next general model in the regression analysis was the

GridSearch CV, which stands for cross-validation. To oversimplify this model we can break it into two pieces first the gridsearch which has to do with all of

the parameters (or restriction the model has) and then cycle through all of the different combinations possible of such parameters. The latter part of the

model is the cross-validation which reduces the variability (or variance from the average). The GridSearch CV is also known as the exhaustive model

because every possible combination is taken, however when the model created scored 100 percent on its test group it lends itself to skepticism. Going

further down the rabbit hole of machine learning our accuracies began to diminish quickly. The K-Nearest Neighbor which groups similar data points

together had an accuracy of seventy-six percent. The last model provided a sampling of several machine learning techniques, only to reconfirm that the

simple Decision Tree Model was had better Principle Component Analysis and Support Vector Machines.

To wrap this random walk through some of the possibilities opened by machine learning models some times occurrence are much easier to predict than we might initially imagine hence complicated models such as XGBoost or RandomForest are unnecessary and even computationally burdensome.