JAYPEE INSTITUTE OF INFORMATION TECHNOLOGY

DEPARTMENT OF COMPUTER SCIENCE AND INFORMATION TECHNOLOGY

CREDIT RISK PREDICTION REPORT

MADE BY :-

Manisha Rathore - 17104033

Akshara Nigam - 17104018

Vatsal Gupta - 17104060

Introduction : The purpose of this project is to build a model that can predict creditworthiness. Credit risk prediction is key to decision-making and transparency.

Decision Tree in Machine Learning is one of the predictive modelling approaches used in statistics, data mining and machine learning. Decision trees are constructed via an algorithmic approach that identifies ways to split a data set based on different conditions. Decision Trees are algorithms that split features into branching paths to arrive at a prediction. Decision Trees are suited to datasets with many features and can outperform Linear or Logistic Regression models.

Algorithms used :

1. **Random forest :-**

   Random forests or random decision forests are an ensemble learning method for classification, regression and other tasks that operate by constructing a multitude of decision trees at training time and outputting the class that is the mode of the classes (classification) or mean prediction of the individual.

   Random Forests are composed of multiple decision trees that take a random sample of the features to form their prediction, and then decide the final classification by consensus vote from all the trees.

   The advantage of a Random Forest model over a simple Decision Tree is that Decision Trees are prone to overfitting. Decision Trees, especially ones that are deep, will form detailed feature branches that fit the training data but don't generalize well.

2. **Decision Tree :-**

A decision tree is a decision support tool that uses a tree-like graph or model of decisions and their possible consequences, including chance event outcomes, resource costs, and utility. It is one way to display an algorithm that only contains conditional control statements. Tree-based learning algorithms are considered to be one of the best and mostly used supervised learning methods. Tree-based methods empower predictive models with high accuracy, stability and ease of interpretation. Unlike linear models, they map non-linear relationships quite well. They are adaptable at solving any kind of problem at hand (classification or regression).

1. **Root Node:** It represents the entire population or sample and this further gets divided into two or more homogeneous sets.
2. **Splitting:** It is a process of dividing a node into two or more sub-nodes.
3. **Decision Node:** When a sub-node splits into further sub-nodes, then it is called the decision node.
4. **Leaf/ Terminal Node:** Nodes do not split is called Leaf or Terminal node.
5. **Pruning:** When we remove sub-nodes of a decision node, this process is called pruning. You can say the opposite process of splitting.

3. Boosting-Based Algorithms :-

Boosting algorithms perform subsequent training by placing weight on data that is hard to classify and less weight on data that is easy to classify. It uses a loss function to measure error and correct for it in the next iteration. Boosted-tree algorithms also penalize models for complexity. The prediction of the final boosted-tree model is the weighted sum of the predictions made by the individual models.

A. **XGBoost :-**

A popular implementation of gradient boosted trees designed for speed and performance. A disadvantage of XGBoost is that it requires data to be stored in memory when run. Since then, newer algorithms have been developed that do not have to process data in memory.

B. **Catboost :-**

The newest of the three algorithms, CatBoost is designed to handle categorical features better. Instead of one-hot encoding features, which causes the curse of dimensionality, CatBoost transforms categorical features into values based on a statistical calculation of its relationship with the target variable. Catboost also divides a given dataset into random permutations and applies ordered boosting on those random permutations.

C. **LightGBM :-**

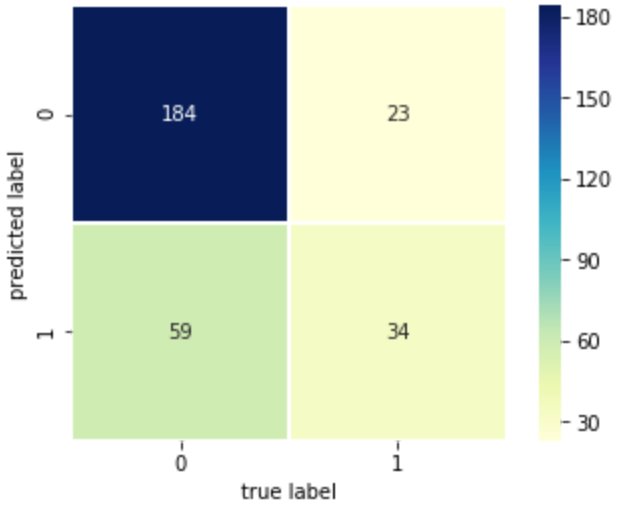
Designed to be a faster implementation of XGBoost with similar accuracy. This algorithm inspects the most informative samples and skips non-informative samples. It also bins sparse features, reducing complexity.

Results

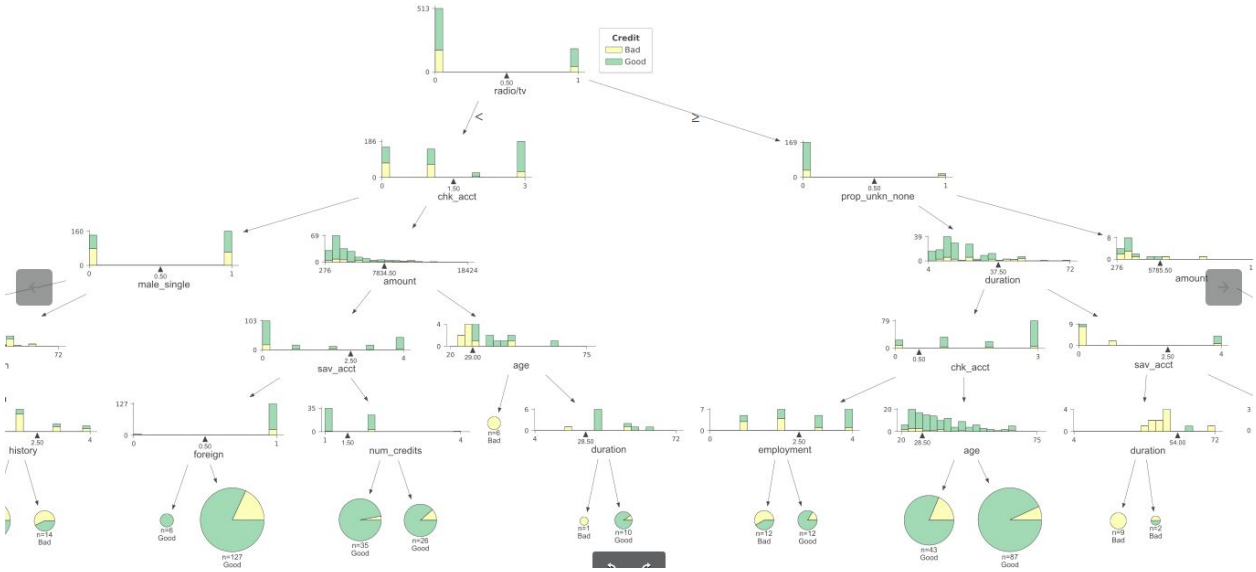
1. **Decision tree model :** The Grid Search found the following best parameters.

   Max depth : 3

   Max feature : 8



Accuracy on test data: 0.727

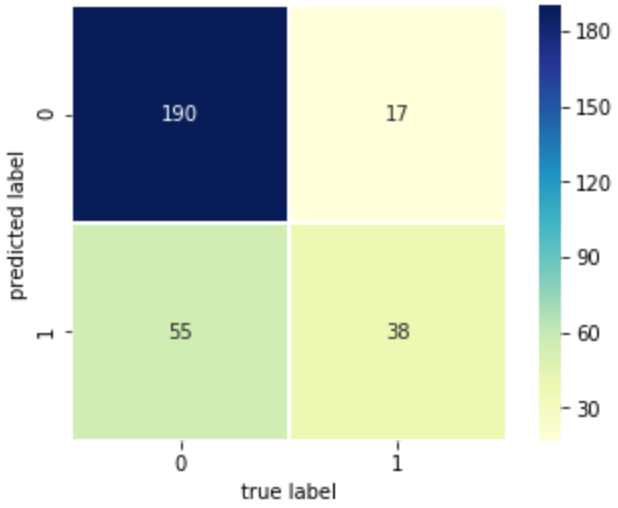


Decision Tree

2. **Random Forest Model :** The Grid Search found the following best parameters.

Bootstrap	:	True
Max depth	:	None
Max features	:	Auto
Min sample leaf	:	1
Min sample split	:	2

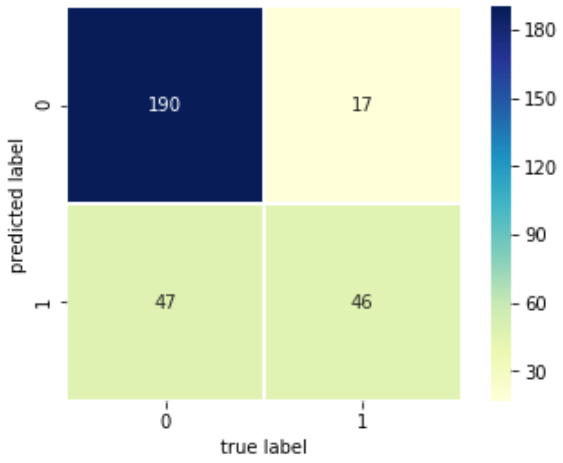
n-estimators : 200



Accuracy on test data: 0.760

By using an ensemble method, the accuracy has improved 4% over the simple Decision Tree model.

3. CatBoost :



Accuracy on test data: 0.786667

4. **LightGBM :**
   Accuracy on test data: 0.7500

5. **XGBoost :**
   Accuracy on test data: 0.766667

   Decision tree model has the highest accuracy on the test data. There was a jump in the accuracy from the Decision Tree model to the ensemble models, which makes sense since ensemble models are designed to perform better through consensus prediction. The Catboost model outperformed the other models with the highest accuracy of 78.7%. This dataset contained many categorical features suited to using the CatBoost algorithm.