Capstone Project
Credit Card Default Prediction (Classification)

(Snehendu Jana)
Data science student
Alma Better

Abstract:

Cardholders are constantly bombarded with solicitations from issuers, seeking to convince them to open new accounts. But what happens when cardholders default on their payments? In this paper, we employ a machine learning approach to predict credit card default. We use a dataset of over 30,000 cardholders, including information on their credit history, revolving balance, and payment history. We find that our models are able to accurately predict credit card default, with an AUC of over 0.87.

The Big data paradigm has revolutionized the banking industry, changing the way financial institutions operate. The aftermath of the past financial crisis has been slowly rectifying and people are now a days better off in terms of job opportunities and financial health The needs for financial services have been increasing drastically over the past years and thus generating huge data in terms of volume, veracity, and variety. Financial institutions have gained in importance and they have been pushed to provide a wide array of services namely; credit facilities, investment, mortgage and retail banking. Therefore, to cope with the phenomenon of budding data, banks are finding ways to leverage their prevailing data sets. The banking industry is gathering massive data from every single transaction of their customers varying from their demographic details to their web history data.

1 . Introduction:

The objective of this case is to predict the probability of default status of credit card holders. This case requires a good understanding of the basic probability concepts as well as the different distribution functions.

A robust model is not only a useful tool for the lending institutions to decide on credit applications. The primary motivation behind prediction is to utilize financial data, for example, amount of the given credit, educational qualification and age, and so forth to foresee the client's need of money or individual credit card data and to decrease loss and vulnerability. Several default prediction models are based on statistical methods, including logistic regression, decision tree classifier, random forest classifier, support vector machine, gradient boosting.

The machine learning model also gives out analysis graphs where we can see that both the classes are not in proportion. An imbalanced dataset is defined by great differences in the distribution of the classes in the dataset. This means that a dataset is biased towards a class in the dataset. If the dataset is biased towards one class, an algorithm trained on the same data will be biased towards the same class.

2 . Methodology:

The default prediction of credit card customers and the analisys cam be broadly divided into 2 working modules. These modules have been describes below;

- **Data Pre-processing**

In order to prepare the data for the experiment, the data is cleaned to remove outliers. It involves removing the null values, renaming the columns, replacing missing values with dummy values, and changing the format of the date and time values so as to utilize them in the algorithm. It also includes removal of outliers.

● Handling class imbalance

Synthetic Minority Oversampling Technique

To solve this problem of imbalance in the dataset is to oversample the examples in the minority class. This can be achieved by simply duplicating examples from the minority class in the training dataset prior to fitting a model. This can balance the class distribution but does not provide any additional information to the model.

An improvement on duplicating examples from the minority class is to synthesize new examples from the minority class. This is a type of data augmentation for tabular data and can be very effective.

Perhaps the most widely used approach to synthesizing new examples is called the Synthetic Minority Oversampling TEchnique, or SMOTE for short. This technique was described by Nitesh Chawla, et al. in their 2002 paper named for the technique titled "SMOTE: Synthetic Minority Over-sampling Technique." SMOTE works by selecting examples that are close in the feature space, drawing a line between the examples in the feature space and drawing a new sample at a point along that line.

● Training and Testing

We used 6 different models.

A – Logistic Regression: Logistic Regression can be used to model the probability of a certain class or event. It is used when the data is linearly separable and the outcome is binary or dichotomous in nature. We used GridSearch to tune the hyper parameters of logistic regression to get the best possible test score.

D – Decision Tree Classifier : The decision trees was also built on the training data in order to improve prediction accuracy .Here also we used GridSearch to tune the hyper parameters of Decision Tree to get the best possible test score.

C – Random Forest Classifier : It builds decision trees on different samples and takes their majority vote for classification and average in case of regression. One of the most important features of the Random Forest Algorithm is that it can handle the data set containing continuous variables as in the case of regression and categorical variables as in the case of classification. It performs better results for classification problems.

D – Support Vector Machine : We use SVMs is because they can find complex relationships between your data without you needing to do a lot of transformations on your own. They typically find more accurate results when compared to other algorithms because of their ability to handle small, complex datasets.

E – Gradient Boosting : The gradient boosting is used to solve classification and regression problems. It is a sequential ensemble learning technique where the performance of the model improves over iterations. This method creates the model in a stage-wise fashion. It infers the model by enabling the optimization of an absolute differentiable loss function. As we add each weak learner, a new model is created that gives a more precise estimation of the response variable.

E – XGBoost : It was used for final prediction of the trip duration in the test dataset. The dataset was very large, as a result for this type of problem XGBoost was applied in which all the attributes were taken and parallel processing of boosting trees executed. Another aspect of XGBoost is that it keeps a nice check between bias and variance which helps in better prediction. The results were interpreted by using GridSearch, the XGBoost hyper parameters.

3 . Dataset of Credit Card clints

We have the dataset containing 30,000 credot card clints. The dataset contains 13 number of features.

Those are :-

- ID - Unique ID of each client
- LIMIT_BAL - Amount of the given credit (NT dollar).
- Gender - Gender of customer. (1 = male; 2 = female)
- Education - Education qualification of customers. (1 = graduate school; 2 = university; 3 = high school; 4 = others)
- Marital Status - Marital status of customer. (1 = married; 2 = single; 3 = others)
- Age - Age of customer in years.
- History of Past Payment - (PAY) Repayment status in September, August, July, June, May and April 2005.
- Amount of Bill Statement - (BILL_AMT) Amount of bill statement in September, August, July, June, May and April 2005.
- Amount of Previous Payment - (PAY_AMT) Amount of previous payment in September, August, July, June, May and April 2005

4 . Steps Involved

• Exploratory Data Analysis

After loading the dataset we performed the method of comparing our target variable that is “IsDefaulter” with other independesnt variables. This process helped us figuring out various aspects and relationships among the target and the independent variables. It gave us a better idea of which feature behaves in which manner compared to the target variable.

• Balancing Imbalanced Data

SYNTHETIC MINORITY OVERSAMPLING TECHNIQUE

It improves the classification of minority classes in imbalanced data. It allows one to oversample the minority class and undersample the majority class. Unlike previous algorithms that oversample the minority class by replication, leading to over-fitting, SMOTE creates synthetic minority data. It over-samples the minority class by taking k (in our case, k = 5) nearest neighbors for a given minority data sample, finding the difference between the features of it and a randomly chosen neighbors, multiplying this difference by a random number between 0 and 1, and adding it to the feature vector.

xnew = xi + (x’i - xi) * a

x’i is one of the K- nearest neighbors of xi , and $\alpha \in [0, 1]$ is a real random number. SMOTE repeats this sampling and perturbation algorithm to create minority data samples according to the amount of over-sampling desired. For instance, over-sampling by 200% creates two new synthetic minority samples by separately perturbing a sample along the vectors of two different nearest neighbors. SMOTE also allows one to undersample the majority class by removing samples until the new majority class is a certain percentage of the original minority class’ sample size. Depending upon the percentage of over and under-sampling, the resulting dataset may have more or fewer samples in the minority class than in the original data. With a slight variation, a similar technique can be used for categorical variables.

Capstone Project
Credit Card Default Prediction (Classification)
(Snehendu Jana)
Data science student
Alma Better

● Standardization of fratures

Our main motive through this step was to scale our data into a uniform format that would allow us to utilize the data in a better way while performing fitting and applying different algorithms to it.

The basic goal was to enforce a level of consistency or uniformity to certain practices or operations within the selected environment.

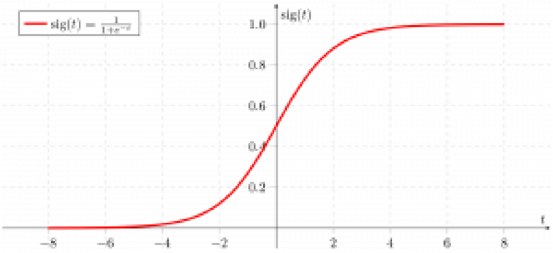
● Fitting Different Models

From modelling er tried various classification algorithms like :

1. **Logistic Regression**
2. **Decision Tree Classifier**
3. **Random Forest Classifier**
4. **Support Vector Machine**
5. **Gradient Boosting**
6. **XG Boosting**

5 . Algorithms

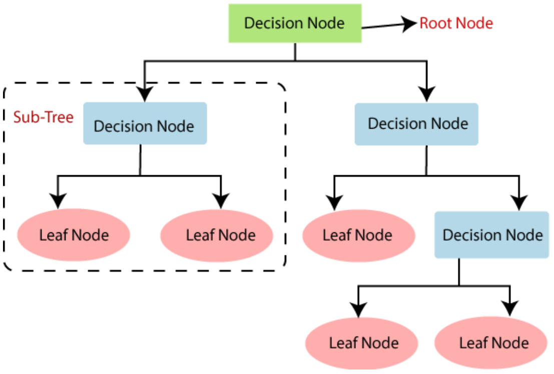
1 . Logistic Regression



Logistic regression is basically a supervised classification algorithm. In a classification problem, the target variable(or output), y, can take only discrete values for a given set of features(or inputs), X. Contrary to popular belief, logistic regression IS a regression model. The model builds a regression model to predict the probability that a given data entry belongs to the category numbered as “1”. Just like Linear regression assumes that the data follows a linear function, Logistic regression models the data using the sigmoid function.

It becomes a classification technique only when a decision threshold is brought into the picture. The setting of the threshold value is a very important aspect of Logistic regression and is dependent on the classification problem itself. The decision for the value of the threshold value is majorly affected by the values of precision and recall. Ideally, we want both precision and recall to be 1, but this seldom is the case.

2. Decision Tree Classifier



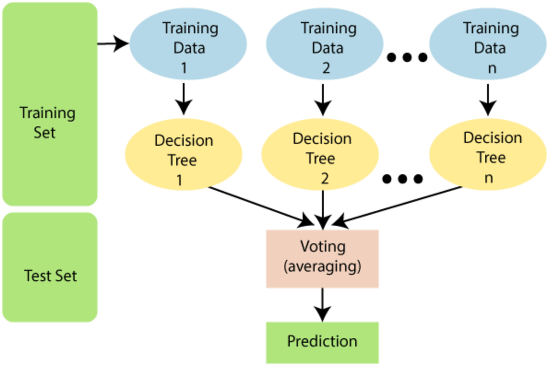
It is a Supervised learning technique that can be used for both classification and Regression problems, but mostly it is preferred for solving Classification problems. It is a tree-structured classifier, where internal nodes represent the features of a dataset, branches represent the decision rules and each leaf node represents the outcome.

In a Decision tree, there are two nodes, which are the Decision Node and Leaf Node. Decision nodes are used to make any decision and have multiple branches, whereas

Leaf nodes are the output of those decisions and do not contain any further branches.

The decisions or the test are performed on the basis of features of the given dataset. It is a graphical representation for getting all the possible solutions to a problem/decision based on given conditions. It is called a decision tree because, similar to a tree, it starts with the root node, which expands on further branches and constructs a tree-like structure. In order to build a tree, we use the CART algorithm, which stands for Classification and Regression Tree algorithm. A decision tree simply asks a question, and based on the answer (Yes/No), it further split the tree into subtrees.

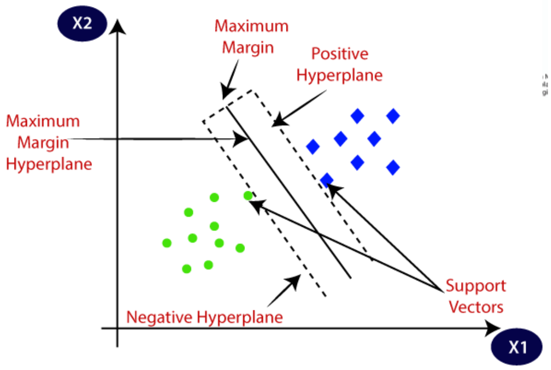
3 . Random Forest Classifier



As the name suggests, "Random Forest is a classifier that contains a number of decision trees on various subsets of the given dataset and takes the average to improve the predictive accuracy of that dataset." Instead of relying on one decision tree, the random forest takes the prediction from each tree and based on the majority votes of predictions, and it predicts the final output. The greater number of trees in the forest leads to higher accuracy and prevents the problem of overfitting.

Random Forest works in two-phase first is to create the random forest by combining N decision tree, and second is to make predictions for each tree created in the first phase.

4 . Support Vector Machine



Support Vector Machine or SVM is one of the most popular Supervised Learning algorithms, which is used for Classification as well as Regression problems. However, primarily, it is used for Classification problems in Machine Learning.

The goal of the SVM algorithm is to create the best line or decision boundary that can segregate n-dimensional space into classes so that we can easily put the new data point in the correct category in the future. This best decision boundary is called a hyperplane.

SVM chooses the extreme points/ vectors that help in creating the hyperplane. These extreme cases are called as support vectors, and hence algorithm is termed as Support Vector Machine. Consider the below diagram in which there are two different categories that are classified using a decision boundary or hyperplane.

5 . Gradient Boosting

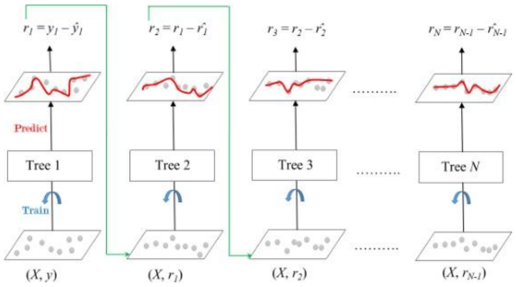
Boosting is one of the popular learning ensemble modeling techniques used to build strong classifiers from various weak classifiers. It starts with building a primary model from available training data sets then it identifies the errors present in the base model.

After identifying the error, a secondary model is built, and further, a third model is introduced in this process. In this way, this process of introducing more models is continued until we get a complete training data set by which model predicts correctly.

Gradient Boosting Machine (GBM) is considered one of the most powerful boosting algorithms.GBM is one of the most popular forward learning ensemble methods in machine learning. It is a powerful technique for building predictive models for regression and classification tasks.

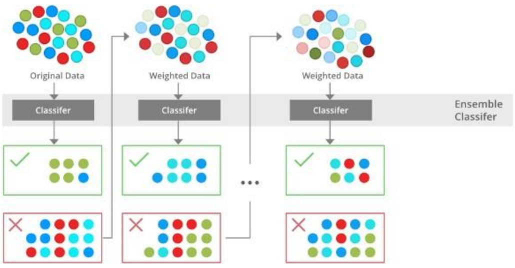
It helps us to get a predictive model in form of an ensemble of weak prediction models such as decision trees. Whenever a decision tree performs as a weak learner then the resulting algorithm is called gradient-boosted trees.

It enables us to combine the predictions from various learner models and build a final predictive model having the correct prediction.



6 . XGBoosting

XGBoost is an implementation of Gradient Boosted decision trees. In this algorithm, decision trees are created in sequential form. Weights play an important role in XGBoost. Weights are assigned to all the independent variables which are then fed into the decision tree which predicts results. The weight of variables predicted wrong by the tree is increased and these variables are then fed to the second decision tree. These individual classifiers/predictors then ensemble to give a strong and more precise model. It can work on regression, classification, ranking, and user-defined prediction problems.



6 . Cross Validation & Hyperparameter Tuning

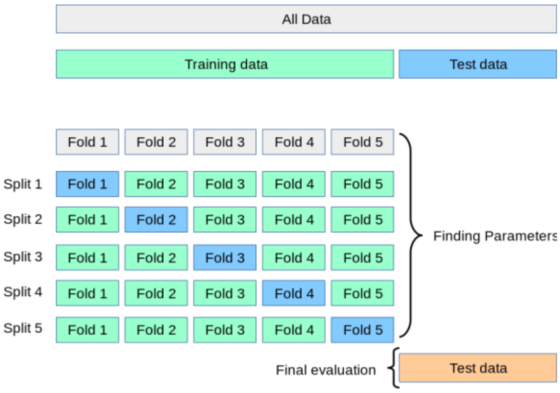
Cross validation and hyperparameter tuning are two tasks that we do together in the data pipeline.

Cross validation is the process of training learners using one set of data and testing it using a different set. We set a default of 5-fold crossvalidation to evalute our results.

Parameter tuning is the process of selecting the values for a model’s parameters that maximize the accuracy of the model.

Capstone Project
Credit Card Default Prediction (Classification)

(Snehendu Jana)
Data science student
Alma Better

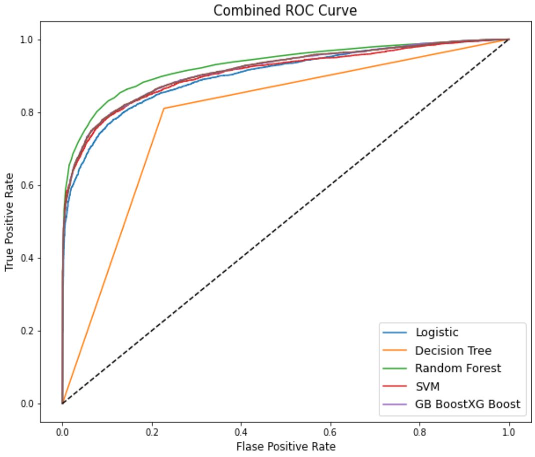


7 . ROC Curve

An ROC curve (receiver operating characteristic curve) is a graph showing the performance of a classification model at all classification thresholds.

An ROC curve plots TPR vs. FPR at different classification thresholds.

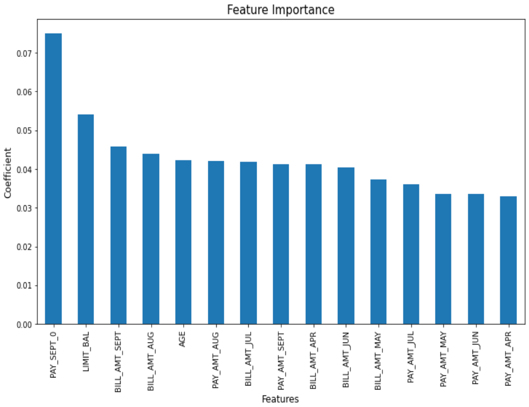
Lowering the classification threshold classifies more items as positive, thus increasing both False Positives and True Positives.



8 . Feature Importance

Feature selection is the process of reducing the number of input variables when developing a predictive model.

It is desirable to reduce the number of input variables to both reduce the computational cost of modeling and, in some cases, to improve the performance of the model.



9 . Conclusion

From all baseline model, Random Forest classifier shows highest test accuracy and F1 score and AUC.

Baseline model of Random Forest and decision tree shows huge difference in train and test accuracy which shows over fitting.

After cross validation and hyper parameter tuning, XG Boost shows highest test accuracy score of 87% and AUC is 0.874.

Cross validation and hyper parameter tuning certainly reduces chances of over fitting and also increases performance of model.