Analytics_Vidhya_Loan_delinquency Problem Solving Approach.

Hello,

1). My Approach:-

"India ML Hiring Hackathon 2019-Loan Delinquency Prediction" is highly imbalanced dataset in which imbalances between both target variables are 99.9% to 0.05 %. S0,classifiers like Logistic Regression, K-nearest neighbour, Naive-Bayes etc. will not work Good here. So I used tree-based

algorithms such as Random Forest Classifier, Adaboost, Catboost, Gradient Boosting, etc. Among all the above-mentioned algorithms Random Forest classifier performs best.

I Tried different models & checked accuracy with classification metrics & then decided to use that model which gives best results.

2). data-preprocessing / feature engineering

To tackle the imbalanced dataset I used 'ADASYN'(AdaptiveSynthetic sampling approach).

ADASYN (Adaptive Synthetic) is an algorithm that generates synthetic data, and its greatest advantages are not copying the same minority data, and generating more data for “harder to learn” examples.

3). Final Model

First used all the algorithms, manually to check which will work more accurately. After this I understand that Random Forest Classifires gives more good results than others, then hyper tuned random forest classifier by varying max depth between [5,20] and min sample split between[1,5] , i know that **Entropy** is the measures of impurity, disorder or uncertainty in a bunch of examples. So used "Entropy" as a criteria.

4).Challenges to handling Imbalanced Datasets

Data Level approach: 1) Resampling Techniques 2)Random Under-Sampling etc

Dealing with imbalanced datasets entails strategies such as improving classification algorithms or balancing classes in the training data (data preprocessing) before providing the data as input to the machine learning algorithm.

feature selection with Genetic Algorithms and Ant Colony Optimization will gives good results.

5). According to Me, Following are Key things a participant must focus on while solving such problems:-

1.) Exploratory Data Analysis

Data analysis is the process of evaluating data using analytical and statistical tools to discover useful information and aid in business decision making. There are a several data analysis methods including data mining, text analytics, business intelligence and data visualization.

-Relationship with Numeric Variables
-Relationship with Categorical Variables

2.) Feature Engineering & Recursive Feature Engineering

Feature engineering is the process of using domain knowledge of the data to create **features** that make machine learning algorithms work. **Feature engineering** is fundamental to the application of machine learning, and is both difficult and expensive.

-For Categorical Variables
-For Numerical Variables

Recursive Feature Elimination (RFE) as its title suggests recursively removes features, builds a model using the remaining attributes and calculates model accuracy. RFE is able to work out the combination of attributes that contribute to the prediction on the target variable (or class).

-RFE (LogisticRegression)
-RFE (RandomForestClassifier)
-RFE (DecisionTreeClassifier)
-RFE (AdaBoostClassifier)

4.) Try All Models

The basic idea of any machine learning model is that it is exposed to a large number of inputs and also supplied the output applicable for them. On analysing more and more data, it tries to figure out the relationship between input and the result.
Try All Following Models:-

- Logistic Regression.
- Decision Tree.
- SVM.
- Naive Bayes.
- kNN.
- K-Means.
- Random Forest.
- Linear Regression.
-

5.) Solving data imbalanced problem

Resampling techniques & Ensemble Methods:-
3 ways:-
-Modify Loss function
-Modify the dataset (resampling)
-Ensemble methods

6) Finally generate test results for sample submission.

- **Confusion Matrix**: A breakdown of predictions into a table showing correct predictions (the diagonal) and the types of incorrect predictions made (what classes incorrect predictions were assigned).
- **Precision**: A measure of a classifiers exactness.
- **Recall**: A measure of a classifiers completeness
- **F1 Score (or F-score)**: A weighted average of precision and recall.