BACKGROUND

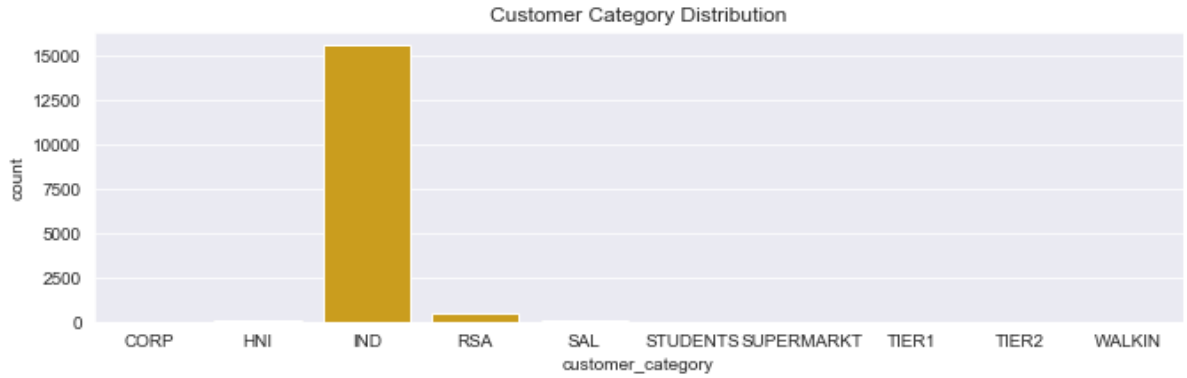
With the enhancement in the banking sector lots of people are applying for bank loans but the bank has its limited assets which it has to grant to limited people only, so finding out to whom the loan can be granted which will be a safer option for the bank is a typical process. So in this project we try to reduce this risk factor behind selecting the safe person so as to save lots of bank efforts and assets. This is done by exploring the Data of the previous records of the people to whom the loan was granted before and on the basis of these records/experiences the model was trained using the machine learning model which gives the most accurate result. The main objective of this project is to predict whether the loan assigned to somebody would be performing or non-performing. The goal of this project is to determine whether a borrower will default on their loan, using both traditional features (salary, level of education, amount in account etc) and non-traditional features that we will engineer are in fact essential, and provides lenders with potentially more information when deciding upon a loan. Analyzing the credit-worthiness of a borrower is an essential step in the loan-making process which has been going on for hundreds of years to varying degrees. Machine learning algorithms can be used to assist institutions in accurately predicting the riskiness of borrowers. We aim to apply a novel approach that uses feature engineering which will hopefully boost predictive power.

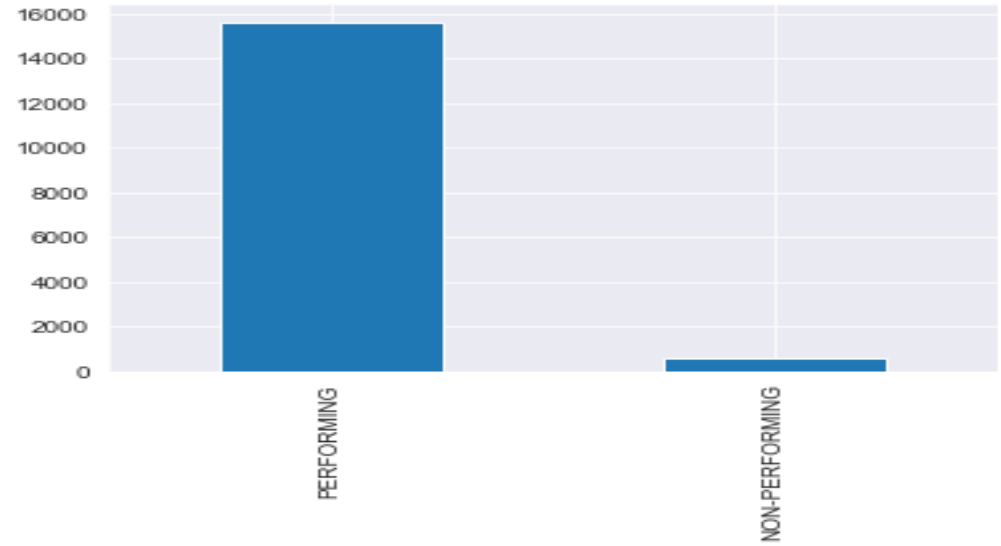
DATA

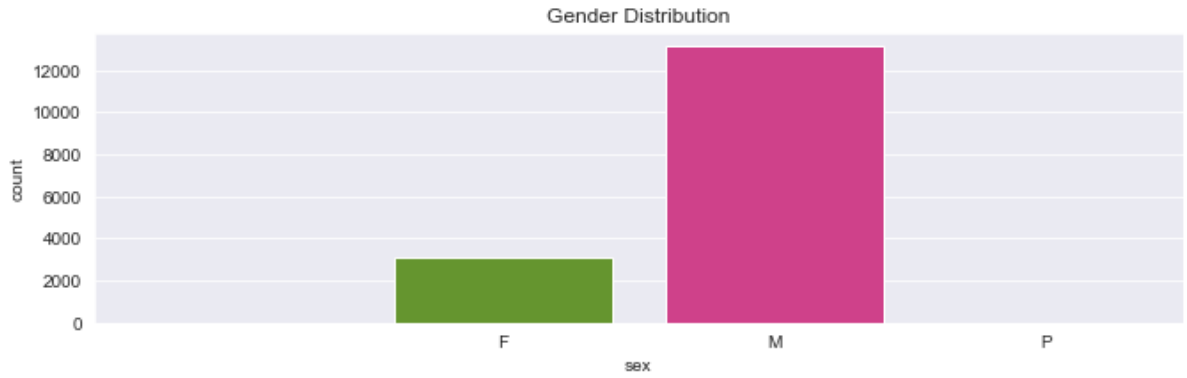
The data set is preprocessed and supplied to machine learning model; on the basis of this data set the model is trained. The data set is split into training and test data set so that after predicting with the train and test data. Every new applicant detail acts as a data set. After the operation of testing, the model predicts whether the new applicant is a fit case for approval of the loan or not based upon the inference it conclude on the basis of the training data sets.

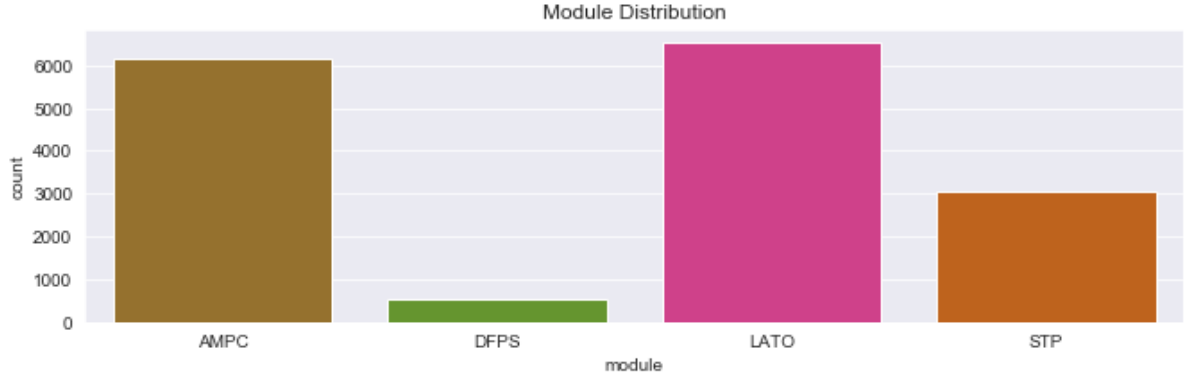
The dataset contains over 18,000 rows, each of them representing either a performing or non-performing loan (binary value) of a financial institution. Performing and non-performing loans are characterized by the 21 columns of data. There is a bias created by imbalanced data, this is because there are more performing than non-performing loans which makes the data set highly imbalanced. After importing the data, we cleaned the variables and removed features with no pertinent information (duplicates for the data, sign with no available entries like ‘NaN’ (Not a Number), for instance). Then, we split the data into two subsets, considering 70% of the data for training the data, and then 30% of these data was used for test purposes. The validation performance permits one to improve the training approach, and we use it to provide prediction performance on the test set. Using the SMOTE technique, we tried to balance the dataset to improve its performance.

VISUALIZATIONS









PREDICTION MODELLING

Logistic Regression

I chose to use a logistic regression model because it is widely used for classification problems. Logistic regression doesn't require linear relationship between dependent and independent variables. It can handle various types of relationships because it applies a non-linear log transformation to the predicted odds ratio

Random Forest Algorithm

I chose to use a Random Forest algorithm because it is impressive in versatility. It can handle binary features, categorical features, and numerical features. Random forest is great with high dimensional data since we are working with subsets of data. Random forest handles outliers by essentially binning them. It is also indifferent to non-linear features.

Decision Tree Algorithm

Decision trees are built using a heuristic called recursive partitioning. This approach is commonly known as 'divide and conquer' because it splits the data into subsets, which are then split repeatedly into even smaller subsets, and so on and so forth until the process stops when the algorithm determines the data within the subsets are sufficiently homogenous, or another stopping criterion has been met.

The most important feature in predicting whether a loan is performing or not performing are:

account_status_a = The status of the bank account

balance – the bank balance

marital_status_M – The married marital status

amount_financed- The amount of money used

applied_amount – The amount of loan applied for

marital_status_S – The single married status

module_LATO – The loan module type called LATO

product_code_PDLP – The product code called PDLP

incr_allowed – The amount of increment allowed in a loan

acy_avl_bal – The total amount in all bank accounts

date_difference – The date difference in the time to repay back a loan

sex_M – The gender of a man

CONCLUSION

While we greatly trimmed our initial data set in order pre-process the data, we felt that the classifiers we trained and used were adequate. Whilst dealing with data imbalance was difficult, using the ensemble modles became a good metric to measure performance alongside accuracy. Out of all the models used to predict whether a loan is performing or not, the Random Forest Classifier and Decision Tree Classifier both perform well with an f1 score of approximately 96%, they have very similar accuracy score of 94.07% and 94.33% respectively

BIAS

It is important to note that the data and the performance of the models are heavily skewed towards performing loans because most of the data involved loans that are performing loans. To develop a more practical system, it is suggested that data involving more non-performing loans be provided to correct this bias.

This model should be handled with care, as it might lead to over fitting when predicting new data. This is because one of the disadvantages of ensemble algorithms is that they are prone to over fitting so when working with new data, close attention should be paid to hyper parameter tuning to ensure the model performs better

FURTHER APPLICATION

An API can be developed which takes in necessary details about the customers and returns a credit score or rating through which the bank can decide if the customer qualifies for a loan or not.