Machine Learning Project -

Prediction of Loan Interest Rate

(Risk based Pricing)

*Business Objective*

A company is in the financial lending business and provides loans to retail customers. It currently uses a process of first collecting certain customer information (attributes), analyzing it manually to decide whether the customer is eligible for a loan or not, and if eligible, telling the customer at what rate of interest rates it can avail the requested loan. Going forward the company wants to employ an automated process which is based upon a Machine Learning model. Not only is this model expected to make the decisioning process faster, but also keep it free from human error and biases.

*Problem Statement*

You have been assigned the task to build for this company a Decision Tree Machine Learning model, where the objective is to predict the loan interest rate category (out of 1, 2 or 3) for a customer. *For more details on the intermediate and final outputs expected, refer to the list of deliverables mentioned in the “Model Building” and “Model Validation” sections below.*

*Data Description*

The dataset provided for this activity consists of 14 features where 13 are independent features and 1 is a target variable. Features in this dataset are described as below :

|  |  |
| --- | --- |
| **Variable Name** | **Description** |
| Loan\_ID | Unique identifier of the Loan |
| Loan\_Amount\_Requested | Loan amount applied |
| Length Employed | Duration |
| Home\_Owner | Ownership of home |
| Annual\_Income | Annual Income of the Loan taker |
| Income\_Verified | Status of income verification |
| Purpose\_Of\_Loan | Category of the purpose for the loan |
| Debt\_To\_Income | The ratio of the representative's total monthly debt repayment divided by self-reported monthly income excluding mortgage |
| Inquiries\_Last\_6Mo | Total number of inquiries in the last 6 months |
| Months\_Since\_Deliquency | The number of days since delinquency. |
| Number\_Open\_Accounts | No of Opened Accounts |
| Total\_Accounts | Total number of credit lines available in representatives credit line |
| Gender | Male/Female |
| Interest\_Rate (Target) | Interest rate categories : 1, 2 or 3 |

*Model Building*

* Show Bi-variate plots (scatter/ bar) of all meaningful variables with the dependent variable
* For the root node, show the Gini value for all the categorical variables?
* Show (using relevant metrics) how much overfitting did you observe with the Raw (unpruned) model?
* Summarize the steps followed to finalize your model - consisting of the below steps (as applicable)
  + Sampling
  + Feature Engineering
  + Performance comparison between Train and Test
  + Use of Cross-validation
* While developing the model, you would have gotten a few candidate models which were not as good as the final model (in terms of performance, multicollinearity, or statistical stability etc.). Show a few of these candidate models and explain their shortcomings
* What approach did you follow in pruning your decision tree model? Which Hyperparameter(s) did you choose and why? Show relevant metrics/output to substantiate your approach
* Show what kind of feature engineering did you apply in your project and why (include in your results what’s applicable from below)
  + Dummy variables
  + Label encoding
  + Any bin-based variable created -what was the significance/rationale of binning
  + Any new derived variables created using the raw variables – For e.g., Ratio based, difference based, % difference based / Rate of change, etc.
* If the provided dataset is unbalanced, what steps did you take to balance it. Also, explain the technique used to oversample/undersample the dataset?
* For a few important independent variables, show how the decision boundary looks like for the finalized model? Hint: At a time, you will need to choose a combination of two independent variables.
* Demonstrate Live how your model will assign class/ or compute the probability for a new data point?
* Provide your understanding of the next steps that the client/ end-user needs to follow to deploy your model at their end. Think about the below lines:
  + Any technical/infrastructure requirements that the client needs to meet?
  + What files do you need to provide them?
  + What kind of data cleaning and preprocessing would the client need to do before using the model?
  + How will the client use your model on new data?
  + How will the client know that the model is performing well on new data points?

Model Validation

* Show your model’s performance on the below metrics (on both train and test samples)
  + Confusion Matrix
  + Classification Report
* For the given business problem which of the below metric(s) did you choose and why? Include in your final output any additional activity performed (and its results) to get to the best values of the below metrics (F1-Score, AUC-ROC curve, AUC-ROC Accuracy).
  + Accuracy
  + Precision
  + Recall