Machine Learning Project -
Predicting Loan Defaults

Business Objective

For a Lending company, it is very important to know whether a customer who availed a loan will pay back the loan or default on it. Thus before investing their money they rely on Predictive Analytics, to figure out which category of customers have a chance of defaulting on future loans. Based on these data driven strategies, these lending companies are able to make smarter decisions in lending money to the right kind of customers.

Problem Statement

You have been assigned the task to build a Decision Tree Machine Learning model for an Online Lending firm “Lending Club”, that predicts the probability of a customer to Default on Loan. *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. Below are the features described.

Index	Variable Name	Meaning
1	credit. policy	it is ‘1’ if the customer meets the credit underwriting criteria of LendingClub.com, and ‘0’ otherwise
2	purpose	The purpose of the loan (takes values "credit_card", "debt_consolidation", "educational", "major_purchase", "small_business", and "all other").
3	int.rate	The interest rate of the loan, proportionate to the amount of riskiness (a rate of 11% would be stored as 0.11). E.g. - Borrowers judged by LendingClub.com to be riskier, are assigned higher interest rates
4	installment	The monthly installments that a borrower owes to the lending company, if the loan is funded.
5	log.annual.inc	The natural log of the self-reported annual income of the borrower.
6	di	The debt-to-income ratio of the borrower (amount of debt divided by annual income).

7	fico	The FICO credit score of the borrower. (FICO is a credit bureau similar to RBI for banks)
8	days.with.cr.line	The number of days the borrower has had a credit line.
9	vol. bal	The borrower's revolving balance (amount unpaid at the end of the credit card billing cycle)
10	vol. util	The borrower's revolving line utilization rate (the amount of the credit line used relative to total credit available).
11	in. last.6 months	The borrower's number of inquiries by creditors in the last 6 months.
12	delinq.2yrs	The number of times the borrower had been 30+ days past due on a payment in the past 2 years.
13	pub.rec	The borrower's number of derogatory public records (bankruptcy filings, tax liens, or judgments).
14	not. fully.paid (Target)	Whether a borrower will fully pay off the loan or not.

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?

- 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 (using relevant metrics) how much overfitting did you observe with the Raw (unpruned) model?

- For a few important independent variables, show what the decision boundary looks like for the finalized model? Hint: At a time, you will need to choose a combination of two independent variables.

- 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, statistical stability etc.). Show a few of these candidate models and explain their shortcomings

- 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 – 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/under-sample the dataset?

- Demonstrate Life 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
    - Concordance test results (This coefficient is used to assess the agreement between estimated values and correct values)
    - Rank ordering test results (Rank ordering is an important measure of model performance and its ability to separate the event from the non-events)

- 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