

Independent Graduate Study Project Proposal (1 Credit) – Harsh Pandya

Overview:

Loan default prediction is a critical application in the financial industry, where lenders and institutions aim to assess the risk associated with providing loans to individuals or businesses. The goal is to identify potential borrowers more likely to default on their loans, allowing lenders to make informed decisions and mitigate financial risks.

Proposed Methodology and Approach

Aim:

This project aims to predict loan defaults using machine learning models preferably Ensemble Techniques and then enhance model interpretability using **Explainable AI techniques**. The project seeks to clearly understand why the models make certain predictions.

Part 1: Data Preparation and Exploration

- Once the dataset is acquired, our next step will be to delve into a comprehensive exploration of its characteristics.
- This will involve a meticulous analysis of data statistics, distributions, and types.
- Understanding these aspects will be essential in gaining insights into the nature and structure of the data, laying the groundwork for subsequent processing.

Part 2: Model Building and Tuning:

- In the imminent phase of model building, we will systematically divide the dataset into distinct training and testing sets, laying the foundation for a thorough evaluation.
- Cutting-edge machine learning algorithms will be employed to unravel intricate patterns and relationships within the data, enhancing our predictive capabilities.
- The integration of Neptune for experiment tracking will empower us to comprehensively monitor and analyze model performance as we embark on this journey.
- The fine-tuning process will deploy state-of-the-art techniques such as Hyperopt and Grid Search, enabling us to meticulously adjust model parameters and optimize predictive accuracy.

Part 3: Explainable AI

- In the forthcoming phases of our project, we will conduct a rigorous evaluation of the overall model performance, employing classification metrics, confusion matrices, and nuanced analyses of recall, precision, and the F1 score. Our exploration will extend to SHAP for intricate feature importance analysis.
- We may also employ techniques such as Anchors to enhance model interpretability if time permits. This forward-looking strategy underscores our commitment to transparency and the continuous refinement of our machine learning models.

Timeline

Weekly Progress and Workload:

Weekly or biweekly updates related to the research progress. Dedicate 6-7 hours (1 credit) per week to research and project-related tasks.

Data Range	Research Phase
Jan 8 - Jan 31, 2024	Data Preparation and Exploration
Feb 1 - Feb 29, 2024	Model Building and Tuning
Mar 1 - Mar 31, 2024	Explainable AI
Apr 1 - May 3, 2024	Explainable AI (continued) + Documentation

Outcome:

- This project will focus on the research's quality, examining the depth and rigor of data analysis, the validity of findings, and alignment with project goals.
- Additionally, the innovative application of machine learning in loan default prediction and the potential real-world impact will be evaluated.