**Project Capstone**

**Project Overview**

Lending Club is seeking the expertise of a data science consultant to perform comprehensive data cleaning, exploratory data analysis (EDA), and predictive modeling on their loan application dataset. The project will also explore the potential for deploying

a real-time scoring application. The primary objective is to prepare the dataset for accurate analysis and modeling, understand the key variables influencing loan approval, and recommend a predictive model for classifying loan applications.

**Task 1: Data Preparation and Cleaning**

**Notebook: DataPreparation&Cleaning.ipynb**

* **Handling missing values:** Missing values were addressed by imputing numerical columns with their respective mean values and filling categorical columns like hardship\_status with "NO\_HARDSHIP". Fully empty rows were also dropped. Also Dropped Unnamed columns.
* **Converting data types:** Certain categorical and numerical fields were processed correctly for consistency, ensuring data integrity for subsequent analyses. E.g. Loan amount was string with $ sign removed the sign and made it numeric value.
* **Removing duplicate records:** Duplicate entries were identified and removed to maintain dataset accuracy.
* **Detecting and handling outliers:** Outliers were detected using Z-score analysis and removed from key columns like last\_pymnt\_amnt to refine model reliability.
* **Standardizing and normalizing data:** Scaling was prepared for future applications where normalization may be necessary, although explicit scaling was not included in this step.
* **Encoding categorical variables:** Job titles were mapped to predefined categories, converting textual information into structured labels.
* **Cleaning and preprocessing string data:** Employee titles were cleaned (lowercased and stripped of whitespace) and flagged for relevant keywords such as "manager."

**Task 2: Exploratory Data Analysis**

**Notebook: ExplorataryDataAnalysis.ipynb**

* **Exploring the Target Variable:** Loan performance was examined across different purposes and grades. Default rates varied significantly across loan categories, revealing trends that could help in risk assessment and interest rate adjustments.
* **Visualizations & Statistical Analysis:** Various plots (bar charts, histograms, box plots) provided insights into loan amounts, default rates, and their distribution among categories. Statistical trends confirmed higher default risks in specific loan purposes and grades.
* **Identifying Predictive Variables:** Loan grade, loan purpose, revolving balance, and utilization rates showed strong predictive relevance for loan performance. Variables with minimal influence were considered for exclusion.
* **Handling Outliers:** High default rates and unusual loan amounts were flagged in lower-grade loans (E, F, G). Some transformations were applied to mitigate the impact of extreme values.
* **Handling Missing Values:** Rows with missing values in revolving balance and utilization were removed to prevent biases in analysis.
* **Examining Feature Relationships:** Correlations between independent variables were evaluated to assess risk exposure, financial stress, and repayment patterns.
* **Class Balance Considerations:** Loan categories showed distinct class distributions, impacting default predictions. Necessary actions were considered to adjust interest rates and financial assistance programs.

**Actionable Insights & Next Steps**

* Adjust interest rates based on risk profiles—lower rates for safer loan grades (A, B, C) and higher rates for riskier ones (E, F, G).
* Implement stricter monitoring for high-risk borrowers, using warning systems for potential defaults.
* Educate borrowers on managing revolving balances and utilization rates to prevent financial distress.
* Further refine feature selection and transformations to optimize predictive modeling.

**Task 3: Modelling**

**Notepad1: Baseline Model: BaselineModel.ipynb**

The process of building and optimizing the loan default prediction model involved several key steps. **First**, extensive data preprocessing ensured the dataset was ready for modeling.

Missing values in critical features such as dti and revol\_util were handled using mean imputation,

while percentage signs in int\_rate and revol\_util were removed to convert these fields to numeric formats.

Categorical variables such as grade and sub\_grade were encoded using ordinal encoding, and additional feature engineering introduced interaction terms and transformations to enhance predictive capability. These steps ensured that the dataset-maintained accuracy, consistency, and scalability for machine learning applications.

For model selection, **logistic regression was chosen as the baseline due to its interpretability and effectiveness in binary classification problems**. The model was trained on a stratified split of the dataset to preserve the distribution of loan defaults. **Accuracy and classification metrics, including precision, recall, F1-score, and ROC-AUC, were used to evaluate its performance.** The initial model achieved high overall accuracy but struggled with recall for defaulted loans, prompting the need for further optimization.

Threshold tuning was employed to minimize financial loss based on misclassification costs, particularly emphasizing the importance of reducing False Negatives, which had a higher penalty. A custom loss function was implemented to balance the trade-off between False Positives and False Negatives, ultimately identifying the optimal classification threshold. Visualization of the loss function across various thresholds helped illustrate the impact of different decision boundaries. The optimized threshold significantly improved recall for defaulted loans while maintaining a reasonable precision balance.

**Comparing different models and evaluation metrics highlighted the improvements introduced by feature engineering and threshold tuning.** Although cross-validation accuracy slightly decreased in the optimized model, the enhanced ROC-AUC score demonstrated its superior ability to differentiate between defaulters and non-defaulters. This refined approach provides a strong foundation for further optimization, ensuring that lending decisions balance profitability with risk mitigation effectively.

**Notepad: Challenger Model: ChallengerModel.ipynb**

The challenger model introduced significant improvements in predictive performance by leveraging advanced feature engineering, ensemble learning, and neural network techniques. The data preprocessing phase carefully addressed missing values, encoded categorical variables, and introduced interaction terms, log transformations, and polynomial features to enhance model interpretability and predictive power. Standardization and encoding ensured the dataset remained consistent across inputs.

The model experimentation involved multiple approaches. The stacking ensemble method combined logistic regression, random forest, and gradient boosting classifiers, yielding a high ROC-AUC score of 0.9786—outperforming the baseline model. Meanwhile, the neural network approach captured deeper non-linear relationships but presented challenges in overfitting, reflected in its ROC-AUC score of 0.7363. Despite its complexity, the ensemble approach demonstrated the best trade-off between precision, recall, and generalizability.

Hyperparameter tuning further optimized model performance, ensuring balanced learning and robust classification results. Through comparative analysis, the ensemble model emerged as the strongest candidate, significantly improving loan default detection accuracy. It provides a scalable, reliable solution for financial risk assessment, reinforcing its suitability for real-world deployment.

**Task 4: Optional Real time scoring application**

**Notepad: Realtimescoringapplication.ipynb**

The real-time loan scoring application is designed to evaluate loan applications instantly by leveraging the trained challenger model. This system processes new loan application data through a structured pipeline, applying preprocessing transformations before generating a default risk prediction. Flask serves as the core framework, enabling seamless API interactions where loan details are submitted via **POST requests** in JSON format.

Key components include **feature engineering**, where categorical and numerical variables undergo encoding and scaling, ensuring compatibility with the machine learning model. The trained **stacking ensemble model** assesses the likelihood of loan defaults, returning both a probability score and classification decision to guide lending decisions.

The application enhances operational efficiency by providing **data-driven risk assessments in real-time**, eliminating manual evaluations while improving consistency. It is designed for **scalability**, making it deployable on **cloud platforms**, integrating **databases**, and expanding to user-friendly interfaces such as web dashboards or mobile apps. With additional security measures and authentication, this system can serve as a reliable solution for financial institutions seeking **instant loan risk assessments**.