elvtr

DATA SCIENCE CONSULTANT BRIEF

DATA SCIENCE IN FINANCE

Client: Lending Club*

Project: Data Cleaning, Exploratory Data Analysis, and Predictive Modeling on Loan Application Dataset

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.

DATASET DESCRIPTION

- **The dataset consists of loan application records stored in a CSV file at the following path:** data/1-raw/lending-club-2007-2020Q3/Loan_status_2007-2020Q3-100ksample.csv
- The dataset contains various attributes such as applicant information, loan details, financial metrics, and application status
- **A data dictionary is provided at the following path:** data/1-raw/lending-club-2007-2020Q3/LCDataDictionary.xlsx

KEY TASKS

01 DATA PREPARATION AND CLEANING

Perform thorough data cleaning on the provided dataset, including but not limited to the following steps:

- Handling missing values (imputation or removal)
- Converting data types to appropriate formats
- Removing duplicate records
- Detecting and handling outliers
- Standardizing and normalizing data
- Encoding categorical variables
- Cleaning and preprocessing string data
- Extracting features from date columns

Perform thorough data clStudents are encouraged to perform additional data cleaning steps beyond those implemented in class (e.g. cleaning of additional columns).eaning on the provided dataset, including but not limited to the following steps:

02 EXPLORATORY DATA ANALYSIS

Conduct an in-depth analysis of the dataset with a focus on the target variable. The analysis should include:

- Exploring the distribution, symmetry, and potential issues with the target variable
- Using visualisation techniques (e.g., histograms, box plots, scatter plots) and statistical analysis to explore relationships between the target variable and independent variables
- Identifying important variables with predictive relevance
- Determining which variables or levels can be excluded
- Identifying variables with outliers and applying transformations if necessary
- Handling missing values and explaining the chosen treatment
- Examining interrelationships between independent variables and considering transformations
- Assessing class balance and addressing any imbalance if needed
- Summarizing insights and plans to leverage the information

03 MODELLING

Recommend and justify a model to predict class membership of loan applications. The modeling phase should include:

- Selecting a baseline model for comparison
- Recommending a challenger model with a detailed justification
- Describing all data preprocessing steps and measurement of accuracy
- Choosing appropriate models and evaluation metrics
- Explaining the choice of models, preprocessing methods, and accuracy metrics

04 OPTIONAL - REAL-TIME SCORING APPLICATION

Build a "real-time" application that can score new loan application observations. The implementation details are at the consultant's discretion.

DELIVERABLES

01 GIT REPOSITORY

- Include all project code with a README file containing a high-level project description
- Example README guide: Make a README

02 REPORT

- Methodology, approach, and model selection rationale
- Advantages and limitations of the chosen model
- Architecture of the final solution
- Considerations on deployment and scalability of the solution - i.e. how will the model be used in BAU by the business?
- Estimated impact/ROI of the project

Note: feel free to cover the above within a jupyter/colab notebook.

TIMELINE

- **Deliverables submission:** 22 October

ADDITIONAL NOTES

- Ensure you justify the choices made throughout your work, and provide comprehensive commentary
- Focus on high-impact approaches to manage the workload effectively

We look forward to your expertise in enhancing the quality and usability of our loan application dataset and providing valuable insights and predictive models.

Best Regards,
Head of Lending

*Note: this is a dummy consultant brief, used for educational purposes only and based on open data sourced from Haggle.