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

Executive Summary

The model helps to identify whether the customer is likely to default on a loan. A dataset of 100,000 customers and a variety of associated information including bank balance, age, and occupation was provided. This was processed and sorted to remove outliers, empty values and any data that may bias the model, or obscure findings. A baseline model was created to evaluate the viability of predictions using standard statistical analysis, and a more advanced 'challenger model' was created to leverage machine learning to classify customers and the likelihood they would default on a loan.

The challenger model (utilising Deep Learning) greatly outperformed the Logistic Regression model and also a 'step up model' utilising a Decision Tree Classifier.

Introduction

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.

Methodology

Data cleaning was initially carried out in order to drop duplicates, remove outliers, eliminate errors and handle missing values.

Exploratory Data Analysis helped in identifying key features which showed correlations with the target variable, identifying the features which would create colinearity (and removing them) and scaling any features with a wide distribution or significant skew.

The baseline model utilised a logistic regression model and tracked FICO and income as well as any hardship payments. It returned the following metrics of performance:

Precision	0.16854952592293726
F1-score	0.26934235976789167
Recall	0.6700080192461909
ROC AUC	0.5920626597851939
Accuracy	0.5340973328536924

The model was good at identifying positive cases but overestimated how many people wouldn't be able to pay back the loans. The low precision however, is not ideal as it may cause a lower amount of loans to be given, reducing revenue for the lending party, leading to customer dissatisfaction (getting rejected on loans they could comfortably pay back) and also leading to poor resource allocation (focusing on strategies to observe and recoup funds from borrowers not in financial hardship).

The 'step up' model utilised a Decision Tree Classifier to determine the likelihood a loan would be defaulted on. It delivered the following metrics:

Precision	0.7407407407407407
F1-Score	0.015866719555573185
Recall	0.00801924619085806
ROC-AUC	0.6313575451700054
Accuracy	0.872501156277301

In this case, the model is a lot more conservative. When it does predict a default, it is usually right. However, it misses a lot of defaults in the process. This is worse for the lender, but better for the customer, as they are less likely to get rejected falsely. This isn't a good compromise though, and this was not suitable as a challenger model.

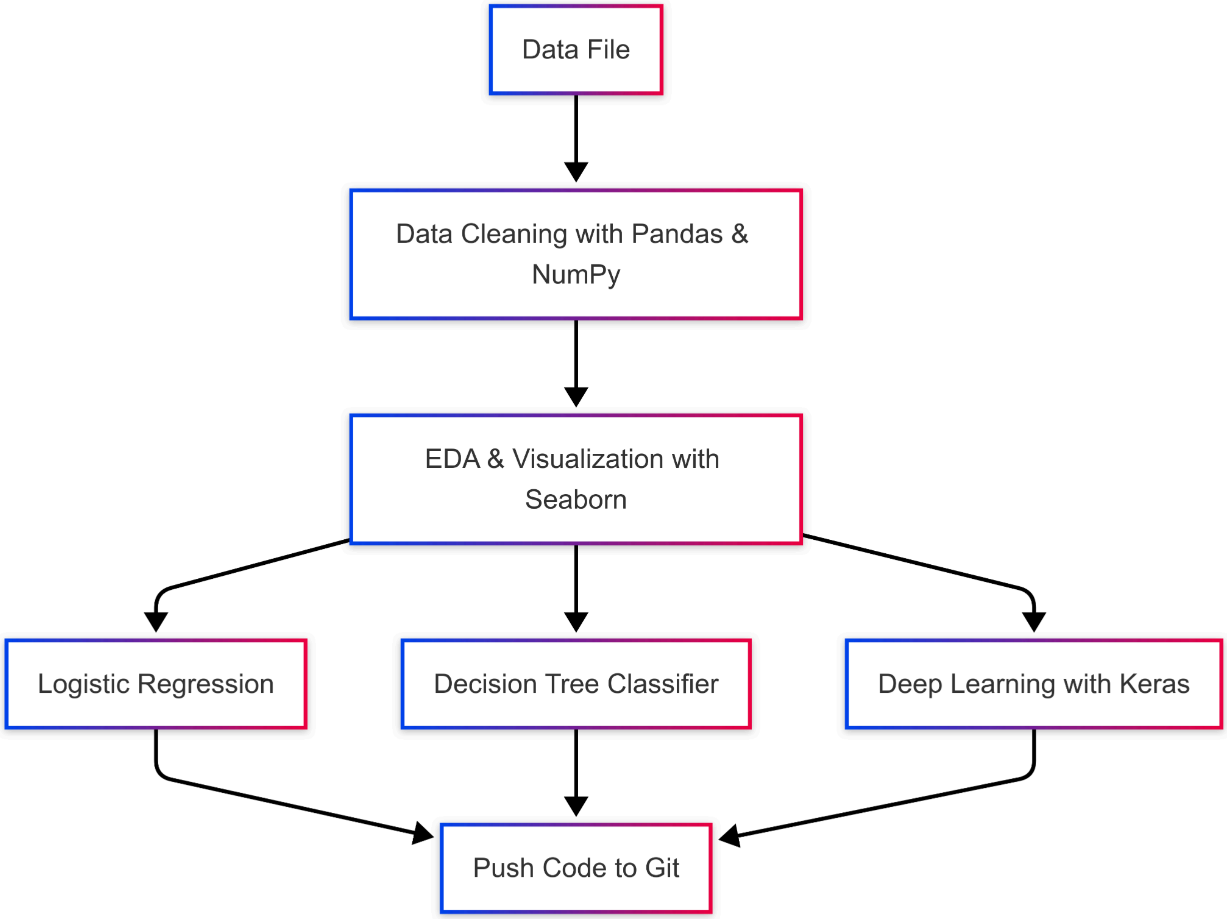
The 'challenger' model utilised Deep Learning, and factored in a lot more features to make its decision. It obtained these results:

Precision	0.618
Recall	0.1494
ROC-AUC	0.8425

Accuracy	0.8813

It now understands risk a lot better, and can determine when a customer will default. There are less likely to be false positives.

Project Architecture



Future Deployment

This model could be called from a web application in the future. Either allowing a user to enter certain metrics about a person and predicting the likelihood of a loan default. They could also set it up so that inputted data is further used to train the model (i.e. from the lending body's records).

The model is not very complex and could be stored in a Git with a main and dev branch, allowing for updates to be made without affecting work in the present. It could be stored on a local server in the short-term but will need to move to a cloud service as the volume of data grows, or if other branches are also participating in the data collection/model utilisation.

Estimated ROI

- Reduced Risk - multiple factors are considered to determine the likelihood of a default. This model is trained on more data than most people will be able to read and recall in the timeframe required to make a decision, and therefore could reduce mistakes made.
- Customer Satisfaction - knowing that personal bias is removed and that the decision is made on an empirical basis may be more acceptable for customers. The model could even give the customer a reason for rejection, which could help them in terms of understanding why they may struggle to get loans, and possibly help them with financial planning.