LendingClub

Data Science Project using PySpark

GitHub Michalina Hulak

Introduction

In this data science project, I delve into the world of lending and finance by leveraging the extensive dataset provided by the Lending Club. Lending Club, as one of the most prominent peer-to-peer lending platforms, has amassed a vast repository of historical data regarding borrowers, their financial profiles, and the outcomes of their loan applications. This treasure trove of information presents us with a unique opportunity to harness the power of data science and machine learning to make informed credit risk assessments.

Goal

My goal is to build a predictive model that can accurately predict whether an individual will repay a loan on time, make late payments, or default on it. The project's objective is to develop a classification model capable of classifying borrowers into these three categories, thereby assisting in credit risk assessment and lending decisions.

Why This Project Matters

This project holds significant importance in the financial and lending domain, offering valuable insights and benefits to multiple stakeholders:

Lending Institutions: Lenders can use the predictive model to make more informed and data-driven lending decisions. By accurately assessing the creditworthiness of applicants, they can minimize default risks and optimize their lending strategies. This leads to better portfolio performance and reduced financial losses.

Borrowers: Borrowers benefit from fair and equitable access to credit opportunities. A robust credit assessment model ensures that deserving individuals are not unfairly denied loans based on inaccurate risk assessments. This promotes financial inclusion and access to capital for personal and business growth.

Risk Management: The project contributes to effective risk management in the lending industry. By accurately classifying borrowers into repayment categories, it helps lenders proactively identify potential defaults and implement risk mitigation strategies.

Financial Stability: Sound lending practices, driven by accurate risk assessment, contribute to the overall stability of the financial sector. Avoiding excessive defaults helps maintain the health of lending institutions, ultimately benefiting the broader economy.

Decision Support: The predictive model can serve as a decision support tool for loan officers and underwriters. It provides them with data-driven insights to supplement their expertise, resulting in more consistent and reliable lending decisions.

Automation and Efficiency: Automation of credit risk assessment through machine learning models streamlines the lending process. This can lead to quicker loan approvals and disbursements, enhancing the overall customer experience.

Results

Logistic Regression

Model	Accuracy (Train)	F1 Score (Train)	Accuracy (Validation)	F1 Score (Validation)	Accuracy (Test)	F1 Score (Test)
LogisticRegression	0.884	0.87	0.886	0.871	0.884	0.869

Random Forest

Model	Accuracy (Train)	F1 Score (Train)	Accuracy (Validation)	F1 Score (Validation)	Accuracy (Test)	F1 Score (Test)
Random Forest	0.87	0.846	0.873	0.849	0.869	0.845

Neural Network

Model	Accuracy (Train)	F1 Score (Train)	Accuracy (Validation)	F1 Score (Validation)	Accuracy (Test)	F1 Score (Test)
Neural Network	0.865	0.845	0.867	0.847	0.865	0.845

Conclusion

In this analysis, I evaluated the performance of three different machine learning models: Logistic Regression, Random Forest, and Neural Network, on both training, validation, and test datasets. The following key observations can be made:

1. **Model Performance:** All three models - Logistic Regression, Random Forest, and Neural Network - achieved competitive performance on all three datasets. They demonstrated relatively high accuracy and F1 scores, indicating their ability to effectively classify and predict outcomes.
2. **Consistency:** Across all three models, we observed a consistent trend in performance. The accuracy and F1 scores on the training, validation, and test datasets were consistently high and very close to each other. This suggests that the models were able to generalize well to unseen data.
3. **Model Selection:** When choosing the best model for a specific task, it's essential to consider factors beyond just performance metrics. Other factors like model complexity, interpretability, and computational resources should also be taken into account. Logistic Regression, for instance, provides a good balance between simplicity and performance, making it a suitable choice when interpretability is crucial.