MGMT 59000-147: Big Data and MLOps

Final Project Presentation

on

Improving Loan Risk Assessment and Approval Process Using Big Data Analytics

Team Members-

- **Tushar Malankar** (tmalanka@purdue.edu)
- **Saquib Hussain** (hussai71@purdue.edu)

Introduction

Why is loan approval a challenge?

- Balancing **risk of defaults** while ensuring smooth approvals.

- Traditional systems **rely on rigid rule-based models** that miss key risk factors.

Our Solution

- Use **Big Data Analytics & Machine Learning** to improve **risk assessment**.
- Optimize loan approvals and **reduce default rates**.
- Predict key financial factors: **Loan default probability, monthly payments, and borrower segmentation**.

Dataset Overview: Loan Default Prediction

Dataset Source

The dataset is from Coursera's Loan Default Prediction Challenge, publicly available on Kaggle.

It consists **of 255,347 unique loan records and 18 columns.**

Key Attributes

Demographics (age, income), Loan Details (amount, term, interest rate), Financial Indicators (DTI, credit lines), and Loan Default Status (0 = No Default, 1 = Default).

The dataset includes demographics, loan details, financial indicators, and loan default status. The goal is to leverage these attributes to predict loan defaults and improve risk assessment.

Data Preparation and Feature Engineering

1 Data Cleaning

Checked for missing values across all columns and confirmed that no missing values were present.

2 Feature Engineering

Created categorical encodings for features like employment type, marital status, and loan purpose.

3 Data Transformation

Standardized numerical features such as income, loan amount, and credit score to ensure better model performance.

Raw financial data requires preprocessing before being used in machine learning models. We focused on preparing the dataset by handling missing values, encoding categorical variables, and creating meaningful features to enhance model performance.

Key Insights from Data Exploration

1 Unique Loans & Missing Values

The dataset contains 255,347 unique loans with no missing values, ensuring data completeness for analysis.

2 Loan Default Distribution

Approximately 11.6% of loans default, while 88.4% are non-default, indicating an imbalanced dataset.

3 Loan Amount & Default

Defaulted loans have a higher average loan amount (≈$144K) than non-defaulted loans (≈$125K), suggesting larger loans may carry higher risks.

Key Insights from Data Exploration (contd.)

Most borrowers have "Very Poor" credit scores (<600), indicating suboptimal credit scores.

Unemployed borrowers have the highest default rate (13.55%), while full-time employees have the lowest (9.46%).

Low-income borrowers (<$30K) have the highest default rate (21.96%), decreasing as income increases.

Exploratory Data Analysis (EDA) Insights

Loan Default Rate

11.6% of loans in the dataset are defaults, while 88.4% are successfully repaid.

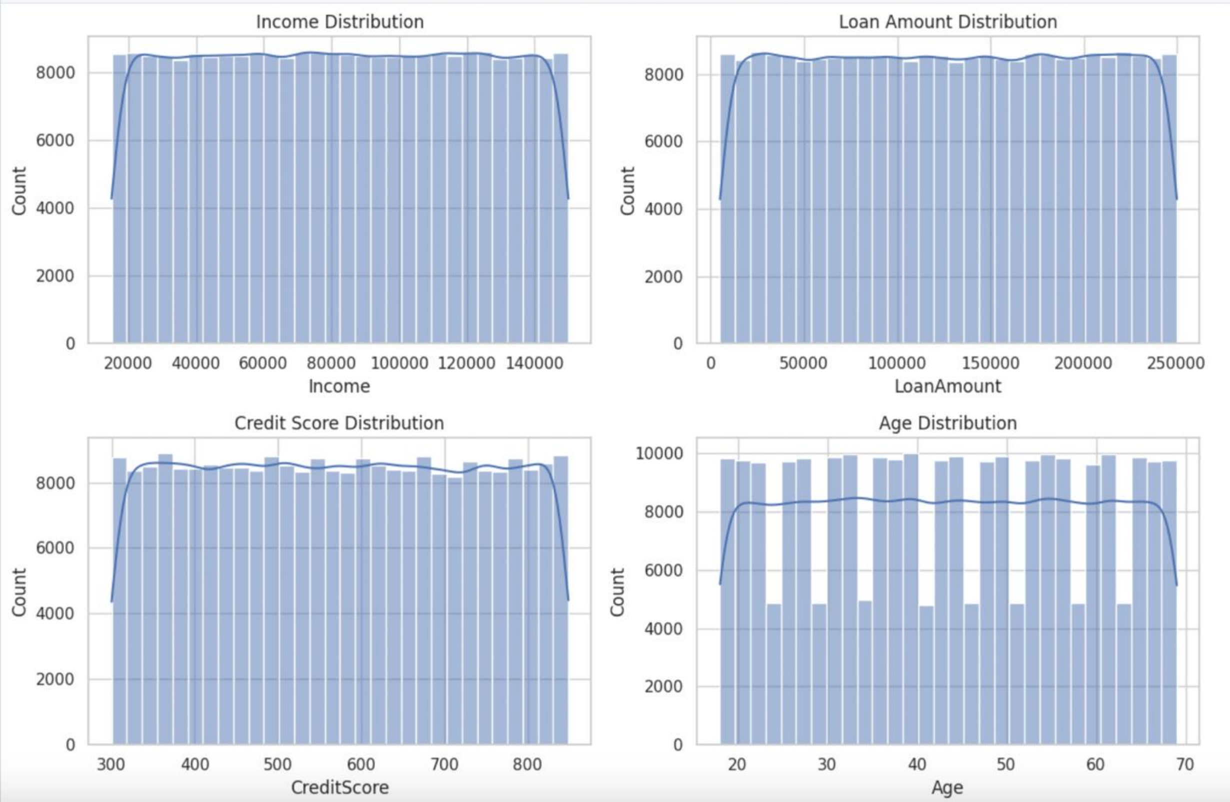
Income vs. Default

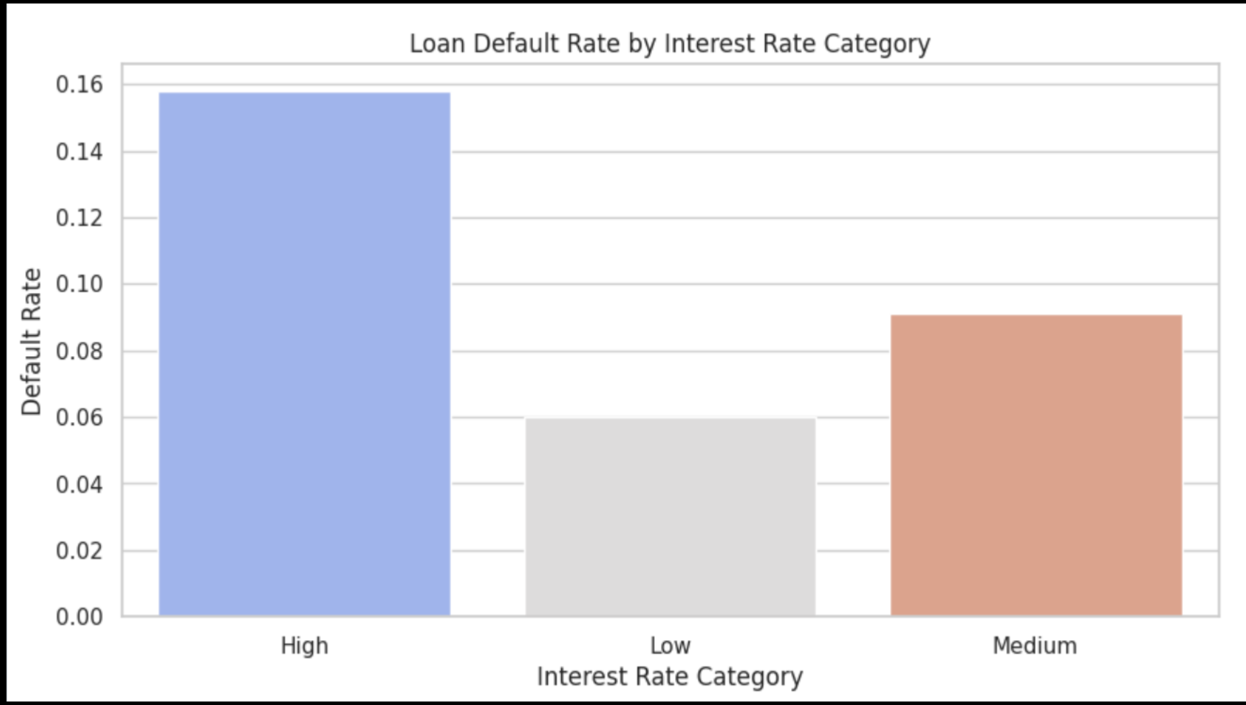
Borrowers with lower income (<$30K) have higher default rates (21.9%), while high-income borrowers (> $100K) have the lowest default rate (9.1%).

Employment Type

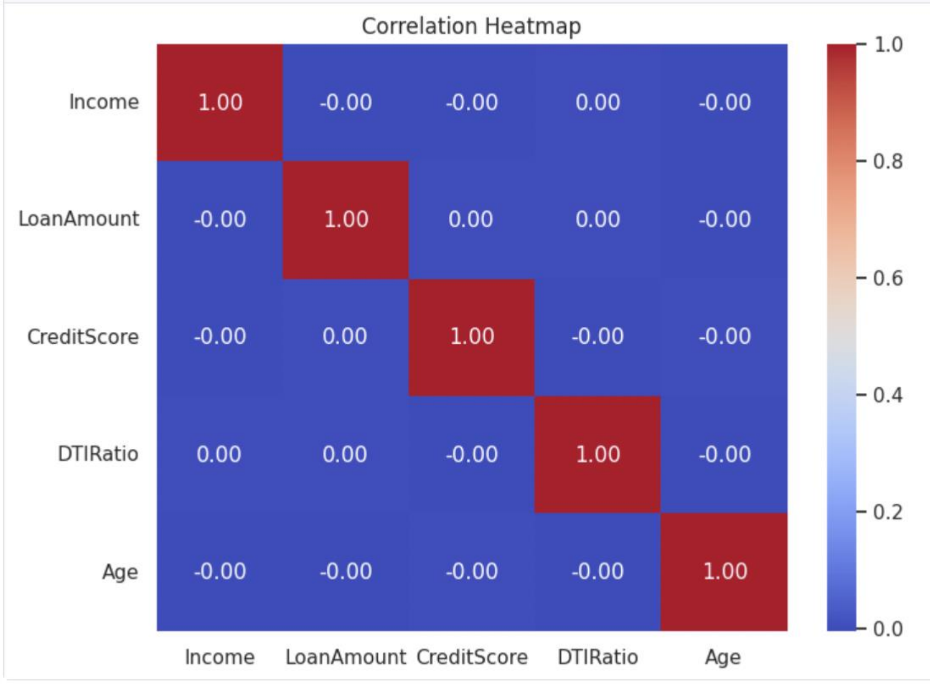
Unemployed individuals had the highest default rate (13.5%), while full-time employees had the lowest default rate (9.4%).

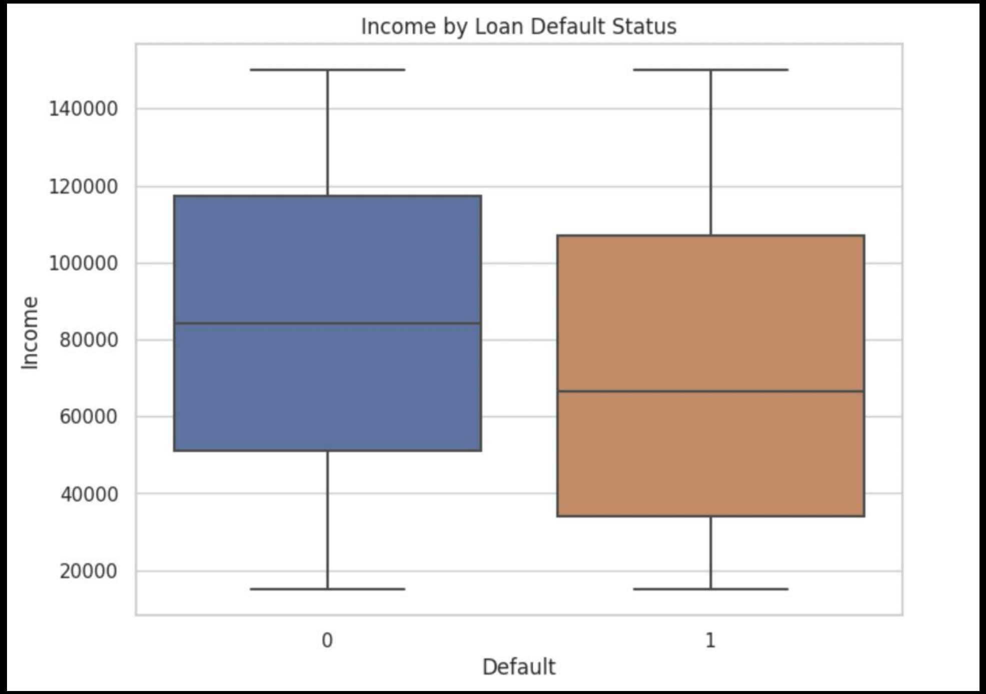
EDA helps in understanding patterns, trends, and relationships within the dataset. We analyzed loan default rates based on borrower characteristics, understood correlations between variables, and identified risk patterns to help financial institutions make data-driven lending decisions.

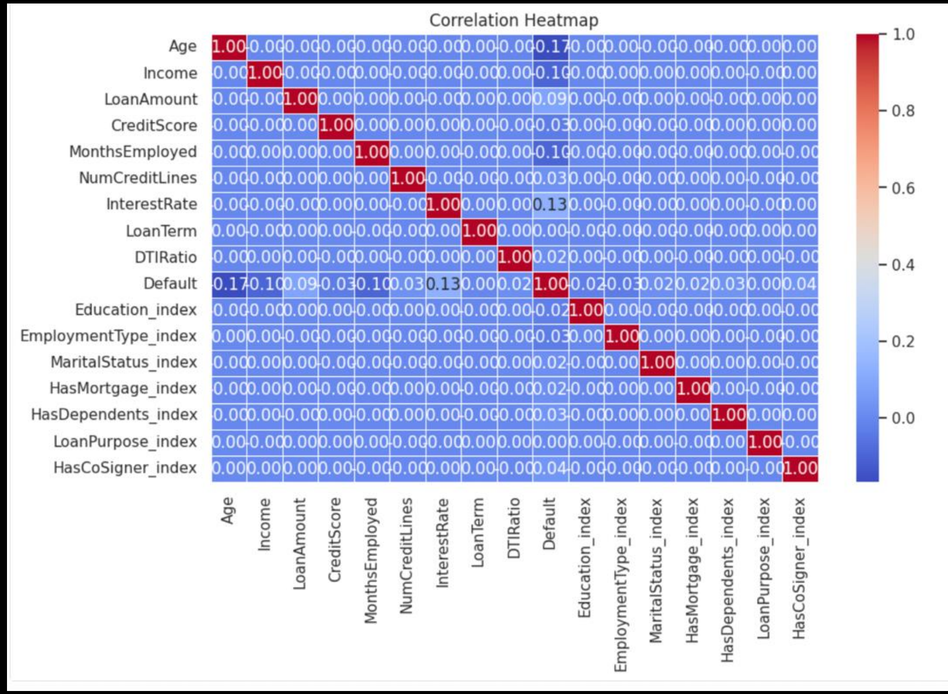


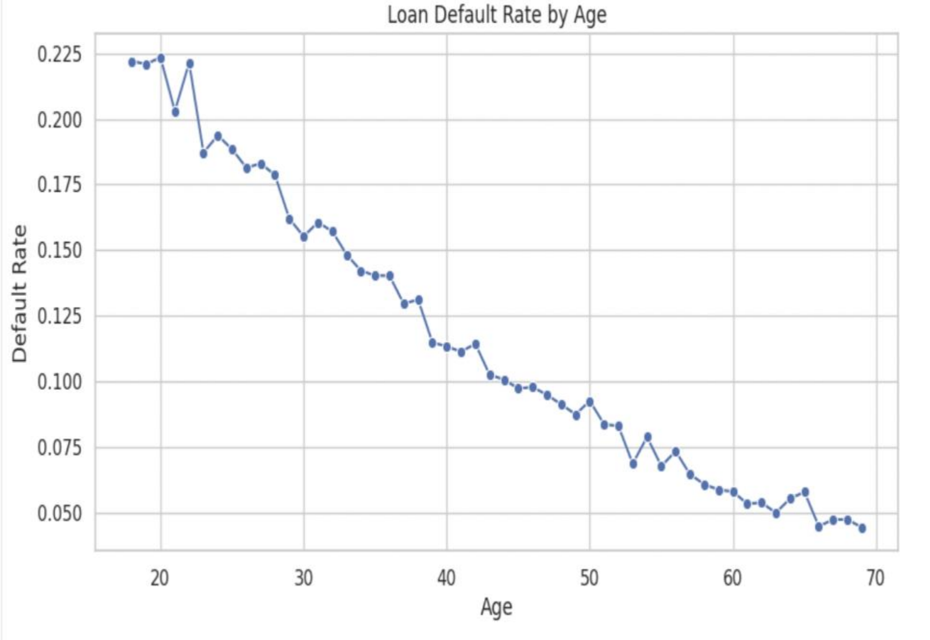


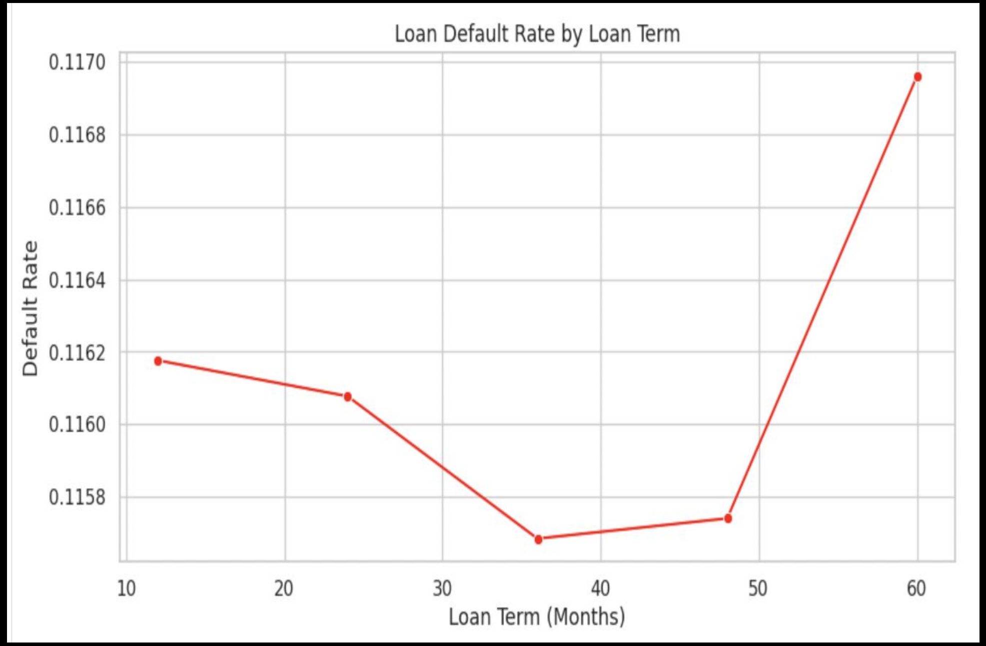
Exploratory Data Analysis

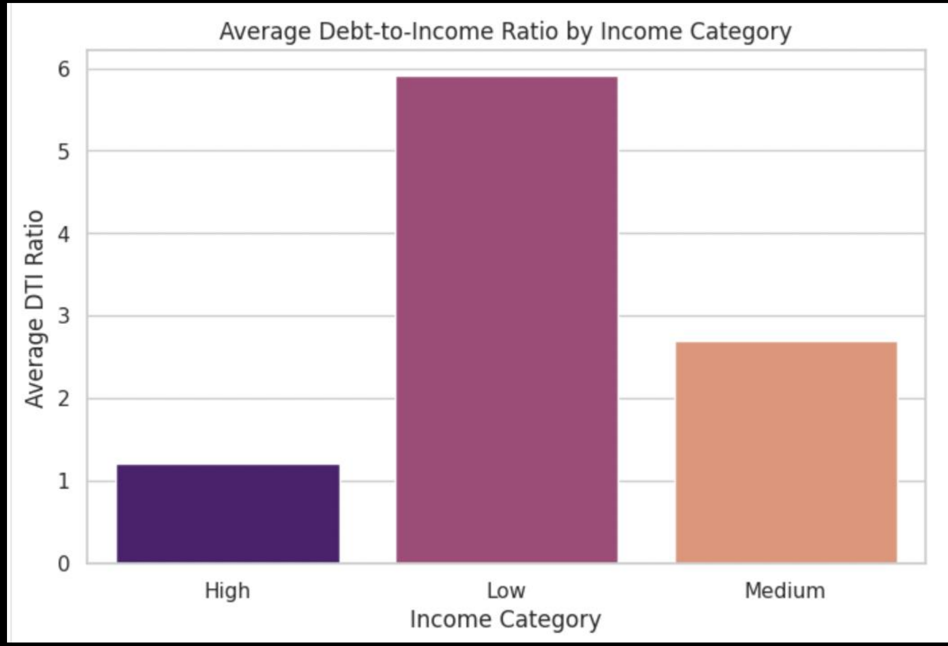


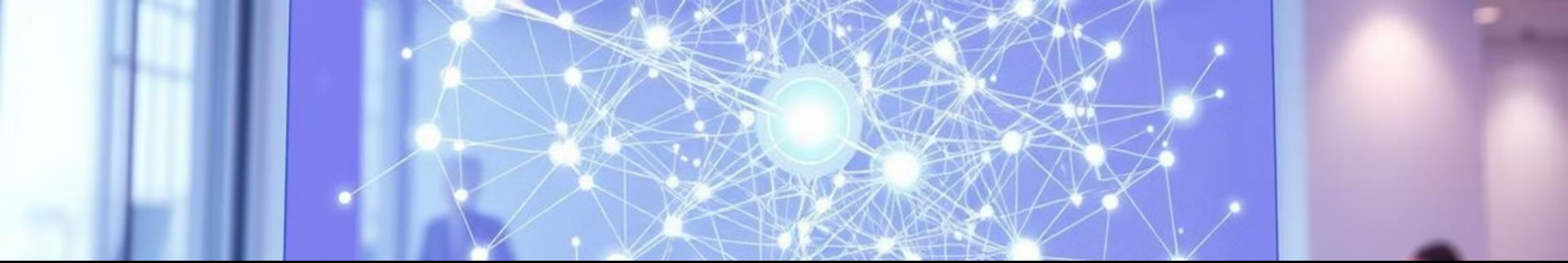












Predictive Insights with Machine Learning Models

1

2

3

Loan Default Prediction

Identifies borrowers at high risk of default, helping financial institutions make informed approval decisions.

Monthly Payment Prediction

Estimates a borrower's expected monthly payment, assisting in affordability assessments and personalized loan structuring.

Customer Segmentation

Groups borrowers based on financial behavior, allowing lenders to tailor loan offerings and risk management strategies.

We built three machine learning models to enhance loan risk assessment and optimize decision-making. These models focus on key financial predictions that can improve lending strategies and customer profiling.

Model 1: Loan Default Prediction Model

Problem Statement

Predict whether a loan applicant will default based on their financial and demographic features.

Model Type

Logistic Regression (Classification)

Evaluation Metrics

AUC, Precision-Recall, Accuracy, F1-score

- AUC: 0.7477574183735693
- Precision-Recall AUC: 0.30283886209441624
- Accuracy: 0.8855047386841794
- Best Regularization Parameter: 0.01
- Best ElasticNet Parameter: 0.0

KEY INSIGHTS

•**Enhance Credit Risk Policies:** Since the model can identify likely defaulters, lenders should **adjust approval criteria based on risk probability**.
•**Segment Borrowers for Personalized Loan Offers:** Customers with **medium risk scores** can be offered **higher interest rates or required to provide collateral**.
•**Improve Loan Recovery Strategies:** Target **potential defaulters** with **pre-emptive financial counseling or structured repayment plans**.

Model 2: Customer Segmentation for Loan Offerings

Problem Statement

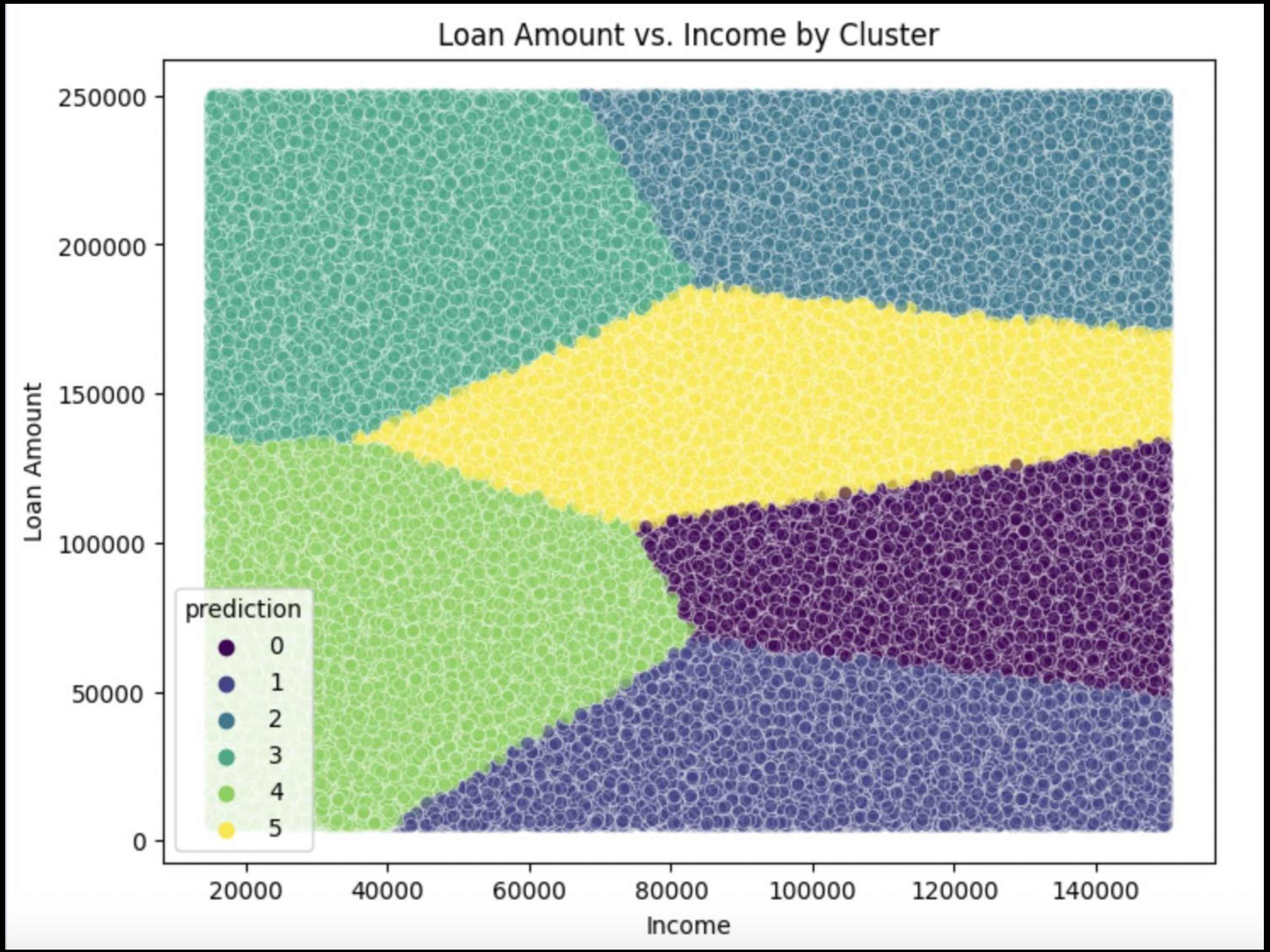
Group loan applicants into different risk categories using unsupervised learning to customize loan offerings.

Model Type

K-Means Clustering

Evaluation Metrics

Within-cluster sum of squares (WCSS), Silhouette Score



KEY INSIGHTS

For High-Risk Clusters (3 & 4):

- •Stricter loan approval policies (higher credit score thresholds).
- •Introduce financial education programs.
- •Offer secured loans (require collateral).

For Low-Risk Clusters (0 & 2):

- •Offer lower interest rates to attract premium borrowers.
- •Provide high-value loan products (business, real estate financing).
- •Cross-sell wealth management & investment products.

Model 3: Monthly Payment Prediction

Problem Statement

Estimate a borrower's expected monthly loan payment based on credit score, income, loan amount, loan term, and debt-to-income ratio.

Model Type

Linear Regression

Evaluation Metrics

RMSE, R^2 Score

KEY INSIGHTS

•**Loan Structuring**: Banks can use this model to tailor **loan terms and interest rates** to match borrower affordability.
•**Risk-Based Pricing**: Borrowers with **higher credit scores and stable income** should be offered **lower interest rates** to encourage responsible borrowing.
•**Targeted Loan Offers**: By identifying customers who **qualify for better loan terms**, lenders can **offer refinancing or pre-approved loans** to drive business growth.

```
✅ RMSE: 2177.118349374987
✅ R² Score: 0.7876820521336133
✅ Selected Features: ['LoanTerm', 'LoanAmount', 'InterestRate', 'Income', 'CreditScore']
✅ Best Optimized Linear Regression Model saved at: dbfs:/mnt/loan_data/final_optimized_lr_model
✅ Monthly Payment Prediction Model Optimization Completed.
```

Additional Model: Early Loan Repayment Prediction

Problem Statement

Predict the probability of a borrower repaying their loan earlier than the scheduled term based on their financial behavior, credit history, and loan details..

Model Type

GBTRegressor

Evaluation Metrics

RMSE, R^2 Score, Mean Absolute Error (MAE)

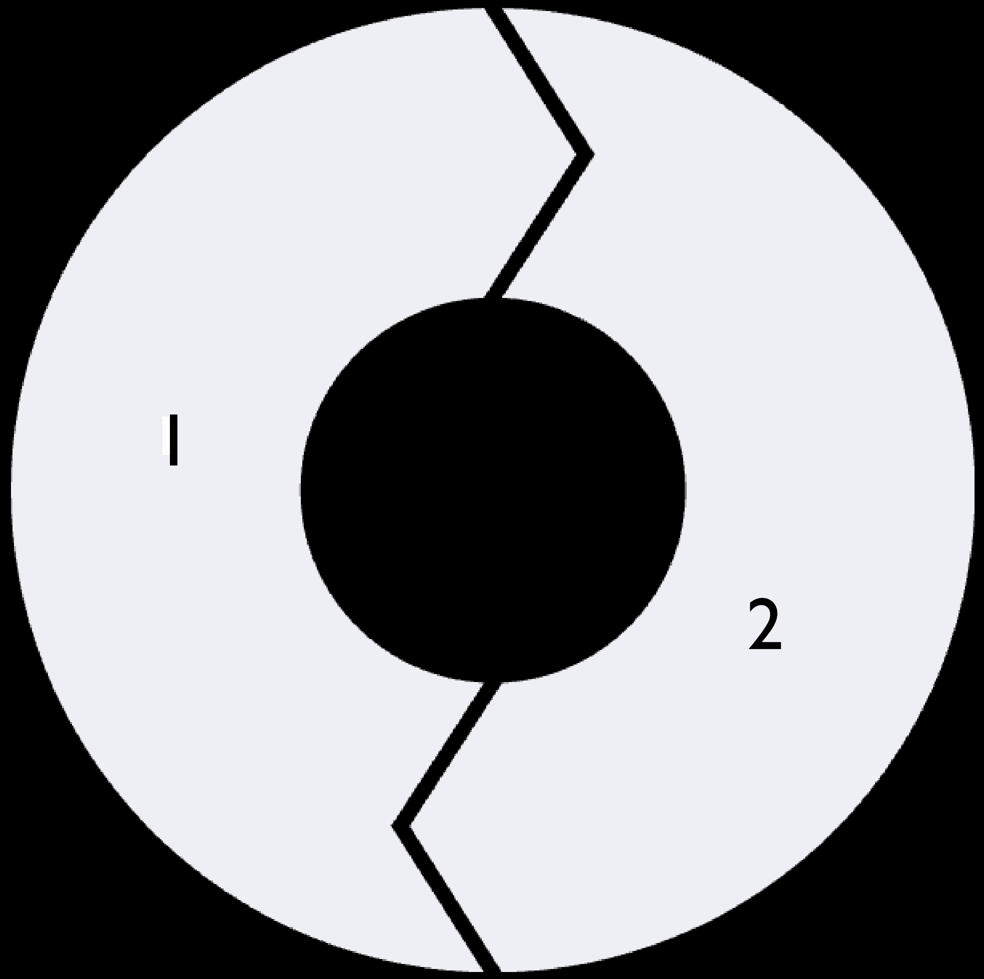
KEY INSIGHTS

•**Personalized Loan Offers**: Financial institutions can **offer better interest rates or flexible repayment terms** to borrowers likely to repay early.
•**Targeted Marketing**: Customers with **high credit scores and stable employment** should be **prioritized for premium loan products**.
•**Risk-Based Lending Adjustments**: Borrowers with **shorter loan terms and strong repayment indicators** may **qualify for lower collateral requirements** or reduced fees.

MLOps Best Practices and Model Performance

Automated Data Processing

Using Spark pipelines to streamline data transformations and feature engineering.

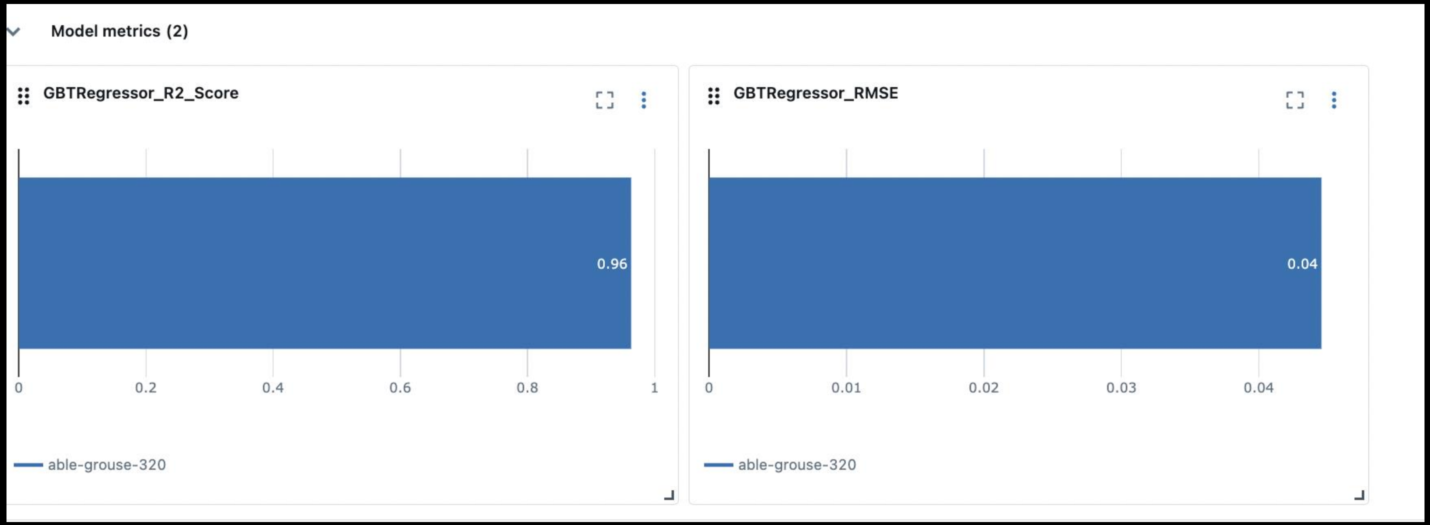


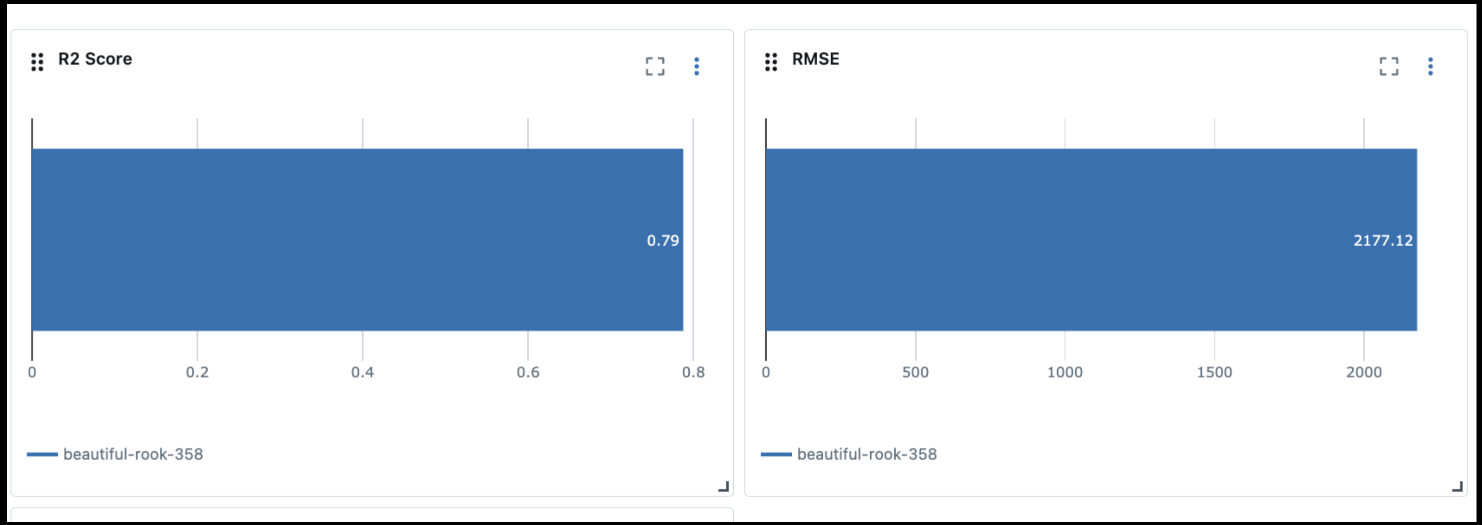
Tracking Model Performance

Using MLflow/Databricks Experiments to log key metrics and hyperparameters.

We integrated MLOps best practices to improve efficiency, reproducibility, and model tracking. Key enhancements included automating data processing using Spark pipelines and tracking model performance using MLflow/Databricks Experiments.

DataBricks Experiment Output





best_model

Path: dbfs:/databricks/mlflow-tracking/2727255611665587/7c4f6e20a0e54139b5710f30f27b2a37/artifacts/best_model

MLflow Model

The code snippets below demonstrate how to make predictions using the logged model. You can also register it to the model registry to version control and deploy as a REST endpoint for real time serving.

Model schema

Input and output schema for your model. Learn more

Name	Type
⊟ Inputs (0)	
No schema. See MLflow docs for how to include input and output schema with your model.	
⊟ Outputs (0)	
No schema. See MLflow docs for how to include input and output schema with your model.	

Validate the model before deployment

Run the following code to validate model inference works on the example input data and logged model dependencies, prior to deploying it to a serving endpoint

```
import mlflow

model_uri = 'runs:/7c4f6e20a0e54139b5710f30f27b2a37/best_model'

# Replace INPUT_EXAMPLE with your own input example to the model
# A valid input example is a data instance suitable for pyfunc prediction
input_data = INPUT_EXAMPLE

# Verify the model with the provided input data using the logged dependencies.
# For more details, refer to:
# https://mlflow.org/docs/latest/models.html#validate-models-before-deployment
mlflow.models.predict(
    model_uri=model_uri,
    input_data=input_data,
    env_manager="uv",
)
```

Make Predictions

Predict on a Pandas DataFrame:

```
import mlflow
logged_model = 'runs:/7c4f6e20a0e54139b5710f30f27b2a37/best_model'
```

Final Insights & Business Impact

📌 **Loan Default Prediction Model**
Successfully identifies high-risk borrowers, reducing bad debt.

📌 **Monthly Payment Prediction Model**
Helps banks customize loans based on affordability.

📌 **Customer Segmentation Model**
Enables risk-based pricing and targeted lending strategies.

📌 **MLOps Best Practices**
Improves model tracking, deployment, and scalability.

Conclusion & Future Improvements

Key Takeaways

✅ Machine Learning enhances loan risk assessment beyond traditional models.
✅ Big Data Analytics provides deep insights into borrower behavior.
✅ MLOps ensures models are scalable, reproducible, and efficient.

Future Scope

- **Integrate alternative credit data** (spending patterns, bank transactions).
- **Explore advanced models** (Deep Learning, Ensemble Methods).
- **Enable real-time loan approvals** using a **streaming architecture**.

Final Thoughts

By continuously improving these models and leveraging new data sources, **financial institutions can strike the right balance between risk mitigation and financial inclusion**.