MGMT 59000-147: Big Data and MLOps

FINAL PROJECT REPORT

Topic- *Improving Loan Risk Assessment and Approval Process Using Big Data Analytics*

Team Members-

Tushar Malankar (tmalanka@purdue.edu)

Saquib Hussain (hussai71@purdue.edu)

Introduction

Approving loans is a key process in the financial industry, helping individuals and businesses access credit. The challenge is to minimize the risk of loan defaults while making the approval process smooth and efficient. Our project uses Big Data Analytics and Machine Learning to improve risk assessment, streamline loan approvals, and reduce defaults.

By using Spark for data processing and machine learning models, we aim to identify patterns in borrower behavior and predict important financial factors like loan default risk, optimal interest rates, and monthly payments.

Traditional loan approval systems often rely on strict rule-based methods, which may not fully capture the complexities of borrower risk. Our approach applies predictive analytics and MLOps best practices to improve decision-making, strengthen risk management, and make the lending process more effective.

Dataset Description

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

The dataset consists of **255,347 unique loan records and 18 columns** with the following key attributes:

- **Demographics**: Age, income, employment type, marital status, etc.
- **Loan Details**: Loan amount, loan term, interest rate, purpose of the loan, and credit score.
- **Financial Indicators**: Debt-to-income ratio (DTI), number of credit lines, presence of a mortgage, and whether the borrower has a co-signer.
- **Loan Default Status**: A binary indicator (0 = No Default, 1 = Default).

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R
1	LoanID	Age	Income	LoanAmount	CreditScore	MonthsEmployed	NumCreditLines	InterestRate	LoanTerm	DTIRatio	Education	EmploymentType	MaritalStatus	HasMortgage	HasDependents	LoanPurpose	HasCoSigner	Default
2	I38PQUQS96	56	85994	50587	520	80	4	15.23	36	0.44	Bachelor's	Full-time	Divorced	Yes	Yes	Other	Yes	0
3	HPSK72WA7R	69	50432	124440	458	15	1	4.81	60	0.68	Master's	Full-time	Married	No	No	Other	Yes	0
4	C1OZ6DPJ8Y	46	84208	129188	451	26	3	21.17	24	0.31	Master's	Unemployed	Divorced	Yes	Yes	Auto	No	1
5	V2KKSFM3UN	32	31713	44799	743	0	3	7.07	24	0.23	High School	Full-time	Married	No	No	Business	No	0
6	EY08JDHTZP	60	20437	9139	633	8	4	6.51	48	0.73	Bachelor's	Unemployed	Divorced	No	Yes	Auto	No	0
7	A9S62RQ7US	25	90298	90448	720	18	2	22.72	24	0.1	High School	Unemployed	Single	Yes	No	Business	Yes	1
8	H8GXPAOS71	38	111188	177025	429	80	1	19.11	12	0.16	Bachelor's	Unemployed	Single	Yes	No	Home	Yes	0
9	0HGZQKJ36W	56	126802	155511	531	67	4	8.15	60	0.43	PhD	Full-time	Married	No	No	Home	Yes	0
10	1R0N3LGNRJ	36	42053	92357	827	83	1	23.94	48	0.2	Bachelor's	Self-employed	Divorced	Yes	No	Education	No	1
11	CM9L1GTT2P	40	132784	228510	480	114	4	9.09	48	0.33	High School	Self-employed	Married	Yes	No	Other	Yes	0
12	IA35XVH6ZO	28	140466	163781	652	94	2	9.08	48	0.23	High School	Unemployed	Married	No	No	Education	No	0
13	Y8UETC3LSG	28	149227	139759	375	56	3	5.84	36	0.8	PhD	Full-time	Divorced	No	No	Education	Yes	1
14	RM6QSRHIYP	41	23265	63527	829	87	4	9.73	60	0.45	Master's	Full-time	Divorced	Yes	No	Auto	Yes	0
15	GX5YQOGROM	53	117550	95744	395	112	4	3.58	24	0.73	High School	Unemployed	Single	No	No	Auto	Yes	0
16	X0BVPZLDC0	57	139699	88143	635	112	4	5.63	48	0.2	Master's	Part-time	Divorced	No	No	Home	No	0
17	O5DM5MPPNA	41	74064	230883	432	31	2	5	60	0.89	Master's	Unemployed	Married	Yes	No	Auto	No	0
18	ZDDRGVTEXS	20	119704	25697	313	49	1	9.63	24	0.28	PhD	Unemployed	Single	Yes	No	Home	No	0
19	9V0FJW7QPB	39	33015	10889	811	106	2	13.56	60	0.66	Master's	Self-employed	Single	Yes	No	Other	No	0
20	O1IKKLC69B	19	40718	78515	319	119	2	14	24	0.17	Bachelor's	Self-employed	Divorced	Yes	No	Education	No	1
21	F7487UU2BF	41	123419	161146	376	65	4	16.96	60	0.39	High School	Self-employed	Single	Yes	No	Other	Yes	0
22	7ASF0IHRIT	61	30142	133714	429	96	1	15.58	12	0.65	PhD	Part-time	Divorced	No	Yes	Business	No	0
23	A22KI1B6SE	47	146113	100621	419	55	1	9.32	12	0.38	Bachelor's	Unemployed	Married	Yes	Yes	Business	No	0
24	1MUSHWD9TW	55	132058	130912	583	48	4	5.82	60	0.47	High School	Unemployed	Married	No	Yes	Business	Yes	0
25	LXK7UEMLK0	19	118989	123300	528	73	3	15.29	36	0.22	PhD	Part-time	Single	Yes	No	Business	Yes	1
26	99ERE1TIB4	38	56848	169818	468	73	1	19.1	24	0.22	Bachelor's	Unemployed	Single	No	No	Education	No	0

Data Ingestion, Processing and Feature Engineering

Raw financial data is often noisy, contains missing values, and requires preprocessing before being used in machine learning models. Before we could find any meaningful insights from the dataset, it is important that we focus on preparing the dataset for modeling by handling missing values, encoding categorical variables, and creating meaningful features.

Steps Taken

- **Data Cleaning**: Checked for missing values across all columns and confirmed that no missing values were present.
- **Feature Engineering**: Created categorical encodings for features like employment type, marital status, and loan purpose.
- **Data Transformation**: Standardized numerical features such as income, loan amount, and credit score to ensure better model performance.
- **Outlier Detection**: Analyzed the distribution of key financial variables to identify and handle potential outliers.

Key Insights

- **No missing values were found**, which means the dataset was already well-prepared for analysis.
- **Loan amount and income distributions are uniform**, indicating that loans are spread across different financial segments.

- Borrowers with **lower credit scores and higher debt-to-income ratios** tend to have a **higher likelihood of loan defaults**.
- **Employment type and loan purpose impact default rates**, with **self-employed** and **unemployed** borrowers showing **higher risks**.

	income_bracket	total_loans	defaults	default_rate
1	Low Income (<30K)	28402	6236	21.95620026758679
2	Mid Income (30K-60K)	56653	7242	12.78308297883607
3	Upper Mid Income (60K-100K)	75727	7566	9.99115242912039
4	High Income (>100K)	94565	8609	9.10379104319780

	LoanPurpose	total_loans	default_loans	default_rate
1	Business	51298	6323	12.32601660883465
2	Auto	50844	6041	11.88144127133978
3	Education	51005	6038	11.83805509263798
4	Other	50914	6002	11.78850610833955
5	Home	51286	5249	10.23476192333190

	EmploymentType	total_loans	default_loans	default_rate
1	Unemployed	63824	8650	13.55289546252194
2	Part-time	64161	7677	11.96521251227381
3	Self-employed	63706	7302	11.46202869431451
4	Full-time	63656	6024	9.46336559004650

Exploratory Data Analysis (EDA)

What is EDA? Why is it Important?

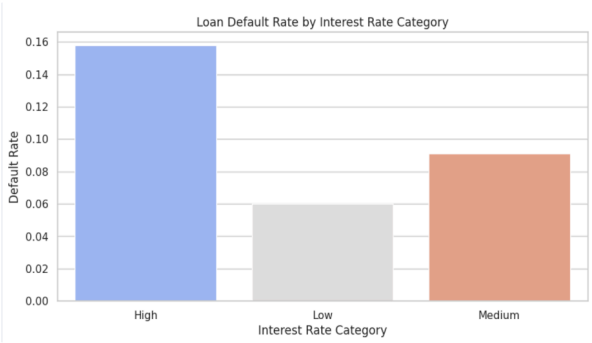
EDA helps in understanding patterns, trends, and relationships within the dataset before applying machine learning models. In the context of our project, we planned on doing EDA to find out basic analytical insights, before we could move on to predictive ones. These are some of the things we tried to look at-

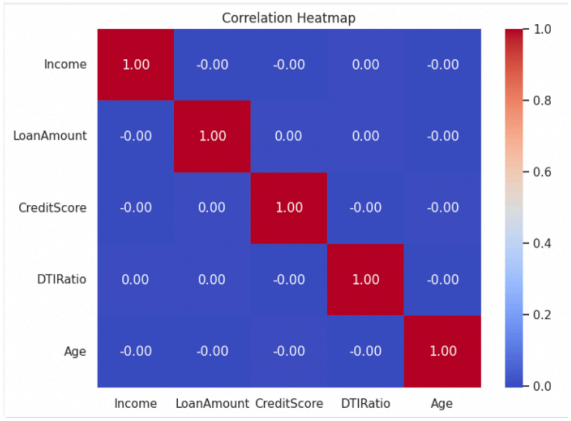
1. **Analyze loan default rates** based on borrower characteristics.
2. **Understand correlations** between variables like income, credit score, and loan amount.

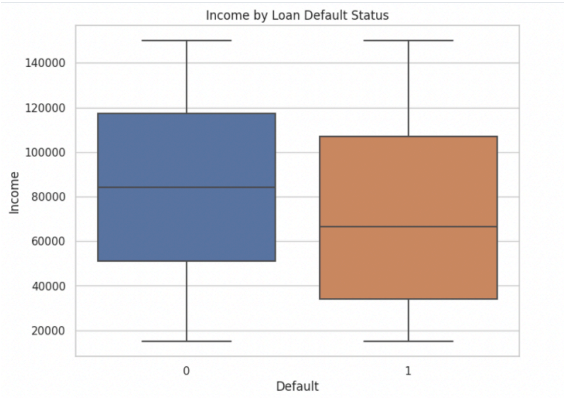
3. **Identify risk patterns** to help financial institutions make data-driven lending decisions.

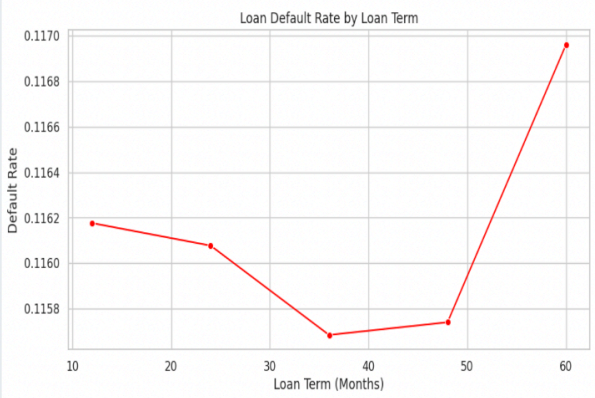
Key Insights from EDA

- Loan Default Rate:
  - **11.6%** of loans in the dataset are defaults, while **88.4%** are successfully repaid.
- Income vs. Loan 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 and Default Rate:
  - Unemployed individuals had the highest default rate (**13.5%**), followed by part-time and self-employed individuals.
  - Full-time employees had the lowest default rate (**9.4%**).
- Credit Score Distribution:
  - A large portion of borrowers had poor credit scores (**<600**), making them high-risk applicants.
- Loan Purpose and Default Rate:
  - Business and auto loans had the highest default rates (**~12.3%**).
  - Home loans had the lowest default rate **(~10.2%**).









Predictive Insights using Machine Learning Models

After conducting Exploratory Data Analysis (EDA) and refining our dataset through feature engineering, 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.

1. **Loan Default Prediction (Classification Model)** – Identifies borrowers at high risk of default, helping financial institutions make informed approval and risk mitigation decisions.
2. **Monthly Payment Prediction (Regression Model)** – Estimates a borrower's expected monthly payment, assisting in affordability assessments and personalized loan structuring.
3. **Customer Segmentation (Clustering Model for Risk Profiling)** – Groups borrowers based on financial behavior, allowing lenders to tailor loan offerings, interest rates, and risk management strategies effectively.

By integrating these models, we aim to reduce default rates, improve loan approval efficiency, and personalize lending strategies based on data-driven insights.

Model Descriptions and Objectives

1. Loan Default Prediction (Logistic Regression)

    - **Objective**: Predict whether a borrower will default on their loan.
    - **Key Features Used**: Credit Score, Income, Loan Amount, Loan Term, Debt-to-Income Ratio.
    - **Performance**: Achieved an AUC of 0.747 and an accuracy of 88.5%, indicating good predictive power.

2. Monthly Payment Prediction (Linear Regression)

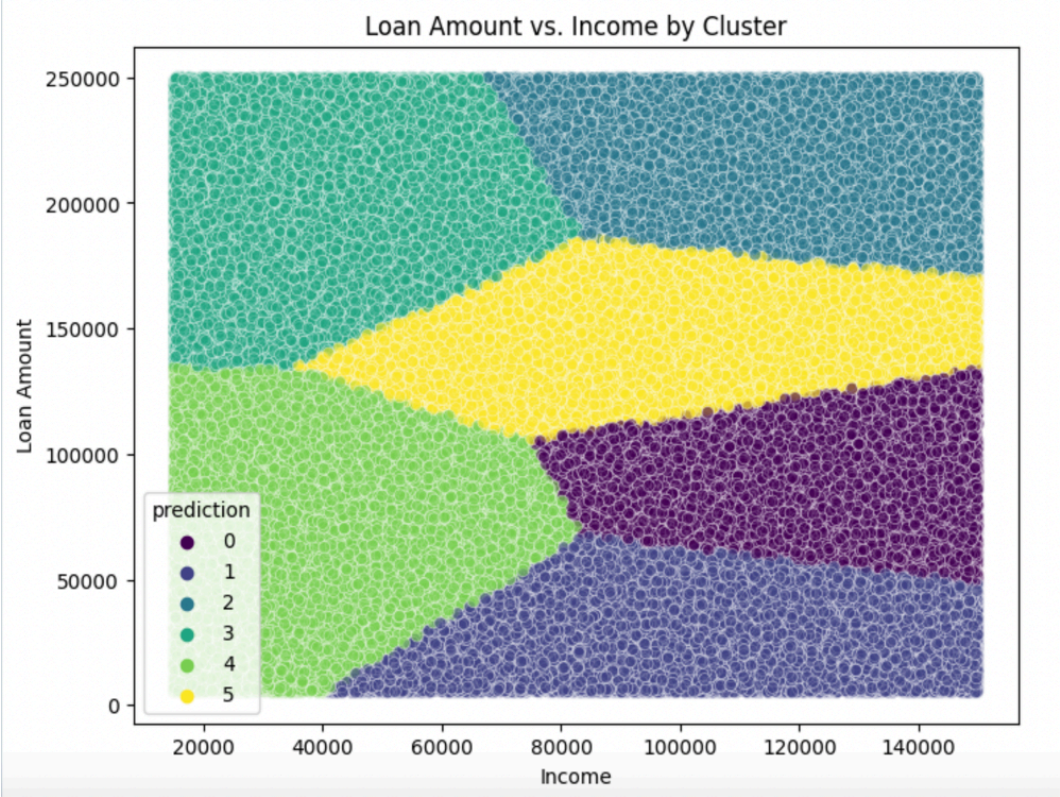
    - **Objective**: Predict the monthly loan payment for a borrower.
    - **Key Features Used**: Loan Amount, Interest Rate, Loan Term, Credit Score, Income.
    - **Performance**: Achieved an R^2 Score of 0.78, meaning the model explains 78% of the variation in monthly payments.

3. Customer Segmentation (K-Means Clustering)

    - **Objective**: Group borrowers into clusters based on financial characteristics and loan risk.
    - **Findings**:
        - High-risk group: Low-income borrowers with large loan amounts and high debt-to-income ratios.
        - Low-risk group: High-income borrowers with moderate loan amounts and high credit scores.

4. Early Loan Repayment Prediction (GBTRegressor)

    - **Objective**: Predict the probability of a borrower repaying their loan earlier than the scheduled term based on their financial behavior, credit history, and loan details. This can help lenders optimize loan structuring and offer early repayment incentives.
    - **Findings**: Features such as **Credit Score, Income, Loan Term,** and **Months Employed** significantly contribute to predicting early repayment likelihood.



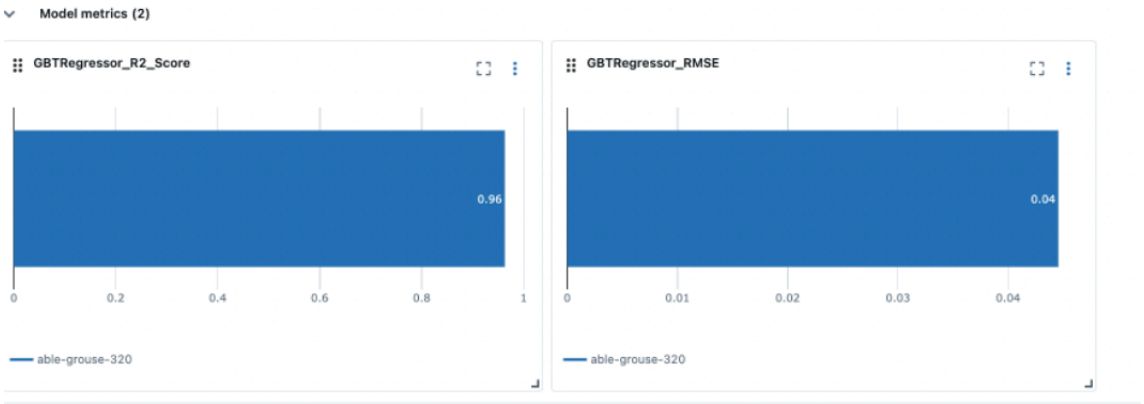
MLOPs Best Practices

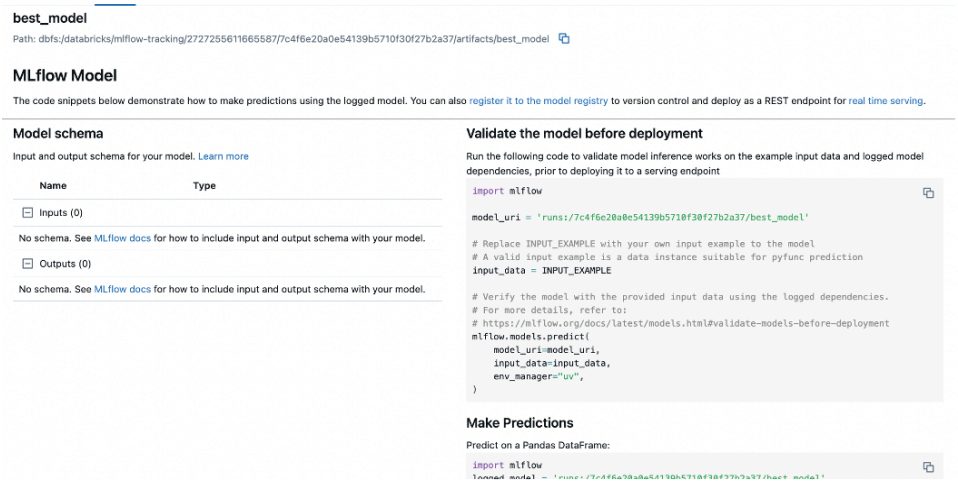
After running all the models and checking their outputs, we re-ran them, but with a focus on integrating MLOps best practices to improve efficiency, reproducibility, and model tracking.

Key enhancements included:

- **Automating Data Processing Using Spark Pipelines** – Spark pipelines streamline the machine learning workflow by automating data transformations, feature engineering, and model training in a structured and scalable manner. This reduces manual intervention, minimizes errors, and ensures consistency across different model runs.
- **Tracking Model Performance Using MLflow/Databricks Experiments** – Keeping track of multiple model runs is crucial for optimizing performance. We created a Databricks Experiment (MLOps Final Project) and then used MLflow to log key metrics, hyperparameters, and model versions, allowing for easy comparison, reproducibility, and deployment of the best-performing models.

By integrating these practices, we enhance the scalability, efficiency, and reliability of our machine learning pipeline, ensuring a smoother transition from model development to deployment.





Insights from the Predictive Models

Loan Default Prediction Model

- The model successfully identifies high-risk borrowers based on credit scores and debt-to-income ratios.
- Can be integrated into automated risk assessment systems for financial institutions.

Monthly Payment Prediction Model

- Helps banks set appropriate interest rates and loan terms to ensure affordability.
- Can be used to tailor financial products for different customer segments.

Customer Segmentation Model

- Allows financial institutions to offer personalized loan products based on borrower risk profiles.
- High-risk borrowers can be provided lower loan amounts or higher interest rates to mitigate risks.

Early Loan Repayment Prediction Model

- **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.

Final Notes

Our project highlights how Big Data and MLOps can transform financial decision-making, making loan approvals smarter and more efficient. By applying machine learning and Spark-based analytics, we were able to:

- Identify key risk factors associated with loan defaults.
- Build predictive models to enhance decision-making for lenders.
- Segment borrowers effectively, allowing for better risk management strategies.

These insights can help financial institutions optimize lending policies, reduce non-performing loans, and improve overall credit accessibility.

Looking ahead, there are several ways this work could be expanded:

- **Integrating alternative credit data** – Including non-traditional indicators like spending patterns, bank transactions, and utility payments could provide a more holistic view of a borrower's creditworthiness.
- **Exploring advanced models** – Techniques like deep learning and ensemble methods could further refine risk predictions and improve accuracy.
- **Enhancing real-time decision-making** – Implementing these models in a streaming architecture could allow lenders to make real-time credit decisions.

By continuously improving these models and incorporating new data sources, financial institutions can strike the right balance between risk mitigation and financial inclusion, ensuring that credit is accessible while keeping defaults under control.