**Capstone\_Project\_1: Loan Analysis & Prediction**

## **Abstract**

This project aims to develop an efficient data pipeline for loan analysis and prediction by integrating big data engineering, real-time streaming, and machine learning. Leveraging Databricks, Kafka, and Power BI, we ensure scalable data processing, insightful visualization, and improved financial decision-making.

## **1. Introduction**

### **1.1 Background**

Traditional loan approval systems often suffer from inefficiencies, leading to inaccurate risk assessment and delayed processing. This project introduces a data-driven approach using machine learning and big data technologies to enhance loan approval decision-making and risk evaluation.

### **1.2 Objectives**

* Develop a scalable data pipeline for loan data processing.
* Implement advanced data engineering techniques to clean and preprocess loan data.
* Train and evaluate multiple machine learning models to predict loan approval probability.
* Visualize key financial insights through interactive Power BI dashboards.
* Deploy real-time predictive models using Apache Kafka for seamless loan processing.

## **2. Literature Review**

To establish a strong theoretical foundation, we explored various research papers and industry reports on loan risk assessment, machine learning applications in finance, and real-time data processing strategies. The study highlights best practices for optimizing financial decision-making through data-driven insights.

## **3. Methodology**

### **3.1 Design & Architecture**

The project follows a structured approach:

* **Data Collection:** Loan-related financial datasets sourced from public and private databases.[[Dataset]](https://www.kaggle.com/datasets/yasserh/loan-default-dataset)
* **Data Engineering:** Data preprocessing, handling missing values, feature selection, and transformation.
* **Machine Learning Models:** Logistic Regression, Random Forest, and Gradient Boosting; fine-tuning the best-performing model.
* **Real-Time Prediction:** Deploying trained ML models using Apache Kafka to enable real-time loan approval.
* **Visualization:** Developing Power BI dashboards for better financial insights.
* **Integration:** Using Databricks for data pipeline scalability.
* **Code Implementation:** [ETL, streaming data processing, and inferencing models in a real-time environment, Data Bricks.](https://colab.research.google.com/drive/1iw1wUE7z32iUCrRBH0oWcb4mVTv9qlIG?authuser=1)

### **3.2 Workflow**

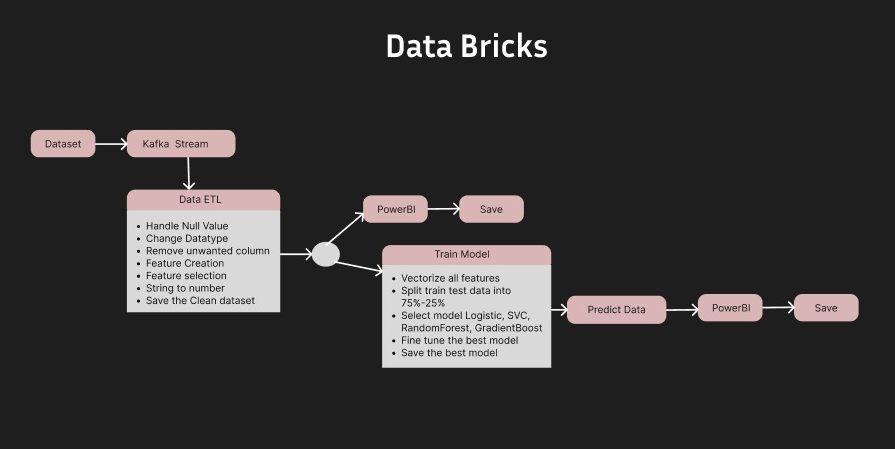


Figure 1: Workflow of the capstone project

## 

## **4. Data Insights & Visualizations**

I created various visualizations using Power BI to gain insights into loan approval trends and risk factors:

* Loan Approval Analysis: Scoreboard representation.
* Occupation-wise Property Value: Pie chart analysis.
* Loan Amount vs. Loan Type: Horizontal bar chart.
* Total Units vs. Security Type: Bar chart.
* Interest Rate and Loan Purpose Analysis: Gantt chart.
* Income vs. Gender: Donut chart for income distribution.
* Credit Score vs. Age & Status: Line plot to understand credit score patterns.
* [Dashboard](https://drive.google.com/drive/u/1/folders/1f_8U5jrQjgIQvd7zsYp2mjtRM6TULyMw)

### **Dashboard Overview**

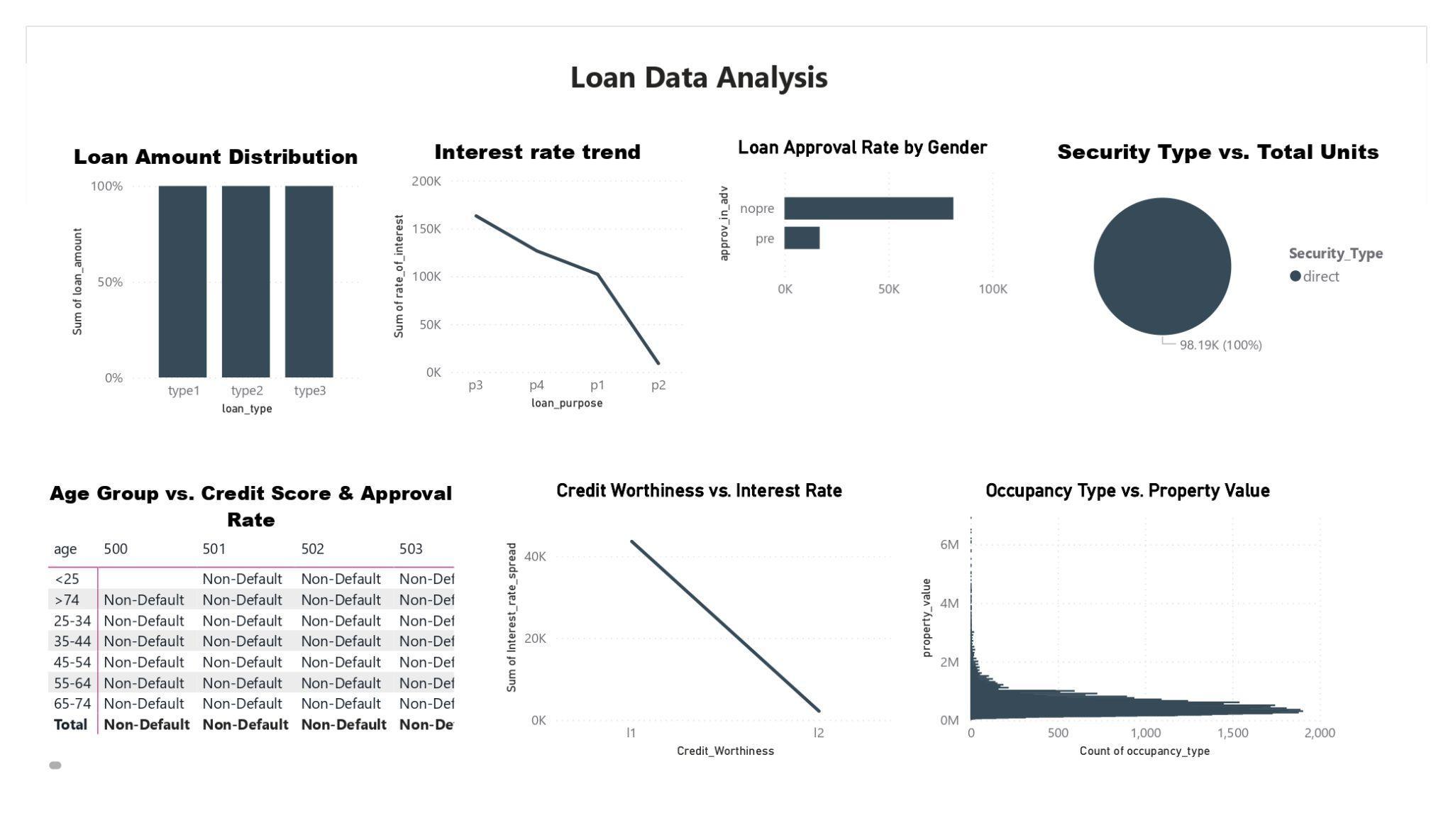


Figure 2: Power BI visualization of loan trends

## 

## 

## **5. Conclusion**

### **5.1 Summary**

This project successfully integrates big data engineering, machine learning, and visualization tools to optimize loan assessment. Through Databricks, Kafka, and Power BI, we ensure real-time predictions and scalable processing for financial institutions.

### **5.2 Future Work**

* Improve model accuracy with additional training datasets.
* Expand data sources to enhance risk assessment models.
* Deploy the solution in a real-world financial environment for further validation.

## **Project Repository & Documentation**

[All in here](https://drive.google.com/drive/u/1/folders/1d2PqoF4HkQRUr8OUY1XCjLO7dgCQ7wmz)

For further details, please refer to the complete documentation in the linked repository.