



# Financial Analytics & Predictive Modeling on Bank Loan Data

A Five-Case Study with PySpark and Databricks

## PRESENTED BY

Sardani Sahithi  
24MBMA64  
MBA GENERAL

# Case 1 - Mortgage Value Prediction

Predicting Potential Mortgage Value

## Method:

A Linear Regression model was trained using features like Income, CCAvg, and Education to predict the Mortgage value.

## Key Finding:

The model confirms a strong link between customer Income and potential mortgage value. However, its primary strength is not in precise forecasting but in identifying high-income prospects, as over 70% of customers have no existing mortgage.

## Business Implication:

This model should be used as a strategic tool to identify and target high-income customers for mortgage marketing campaigns, rather than for underwriting specific loan amounts.

# Case 2&3: Understanding the Customer: Segmentation & Risk Prediction

## Method:

- **Unsupervised Clustering (K-Means):** Deployed to segment the customer base into three personas based on their financial profiles for strategic targeting.
- Supervised Classification (Logistic Regression): Trained to build a predictive model with ~96% accuracy for tactical screening of individual loan applicants.

## Key Finding:

- Three Actionable Segments Emerged: The data clearly shows three personas:
  - "High-Value Prospects": High income and high credit card usage.
  - "Developing Middle-Market": Moderate income and spending.
  - "Stable & Cautious": Low income and minimal credit usage.
- Prediction is Highly Accurate: The risk model can predict loan acceptance with ~96% accuracy, making it a reliable tool for campaign targeting.

# Case 4 - Simulating Market Changes

## Simulating the Impact of Interest Rate Hikes

### Method:

Performed a simulation by using the trained risk model (from Case 2) on data where the CCAvg feature was artificially increased to model the impact of rising interest rates.

### Key Finding:

The simulation revealed that the portfolio is highly sensitive to market changes. A hypothetical 50% increase in a customer's credit card burden (simulating a rate hike) would cause the average probability of loan acceptance to drop by nearly 20%.

### Business Implication:

This model provides a quantitative tool for risk management. The bank can use it to "stress-test" the portfolio and develop proactive strategies to mitigate the impact of economic downturns on loan origination.

# Case 5 - Drivers of Financial Stability

## Identifying the Key Drivers of Financial Stability

### Method:

Trained a Decision Tree Classifier to predict which customers have a CD Account (a proxy for financial stability) and to identify the most influential predictive factors.

### Key Finding:

A Decision Tree model identified Income as the single most powerful predictor of financial stability (defined by having a CD Account). Whether a customer already holds a Personal Loan was also a significant factor.

### Business Implication:

This confirms that Income should be the primary factor in identifying low-risk customers for cross-selling opportunities. The model provides clear, data-driven rules for targeting the most stable segments of our customer base.

# Project Summary

## Key Business Takeaways:

- Prediction is Highly Viable: Customer behavior, such as loan acceptance and financial stability, can be predicted with high accuracy.
- Segmentation Unlocks Value: Our customer base is not monolithic; it consists of distinct personas that can be targeted for tailored marketing and product offerings.
- Simulations Quantify Risk: We can now model the financial impact of external market changes, moving from reactive to proactive risk management.

## Technology Stack Used

- Platform: Databricks Community Edition
- Core Engine: Apache Spark (via PySpark)
- Libraries:
  - Machine Learning: Spark MLlib
  - Data Manipulation: Spark SQL, Pandas
  - Visualization: Matplotlib, Seaborn, Databricks