# Leveraging Machine Learning to Predict Loan Defaults

## Phase 3: Feature Engineering and Hyperparameter Tuning


Group 4

Maria P. Aroca, Cassie Cagwin, Lexi Colwell, Nasheed Jafri

# Project Overview

- **Purpose:** The project aims to help lending institutions improve loan approval processes by predicting an applicant's ability to repay loans using a combination of Telco and Transactional data.
- **Problem:** Many loan applications are rejected due to insufficient credit history, forcing applicants to turn to predatory lenders. This project seeks to reduce unjust rejections while minimizing the risk of defaults.
- **Goal:** Build a predictive system to identify potential defaulters using applicant data.
- **Data**: The dataset from the Home Credit Default Risk project comprises a rich and diverse collection of data sources:
  - **Application Data:** Demographic and financial details for current loan applications
  - **Credit Bureau Data:** Records of clients' past credits and monthly balances (other sources)
  - **Credit Card Balances:** Records of monthly balances for previous Home Credit credit cards
  - **POS & Cash Loans:** Monthly snapshots of repayment histories for POS and cash loans
  - **Installment Payments:** Records of repayment histories, including missed installments
  - **Previous Applications:** Records of past loan applications with Home Credit

# Phase 3 Work: Key Tasks Completed

### Additional Feature Engineering

We revised our feature engineering, particularly for the bureau_balance and credit_card_balance data sets, to add additional features for our models.

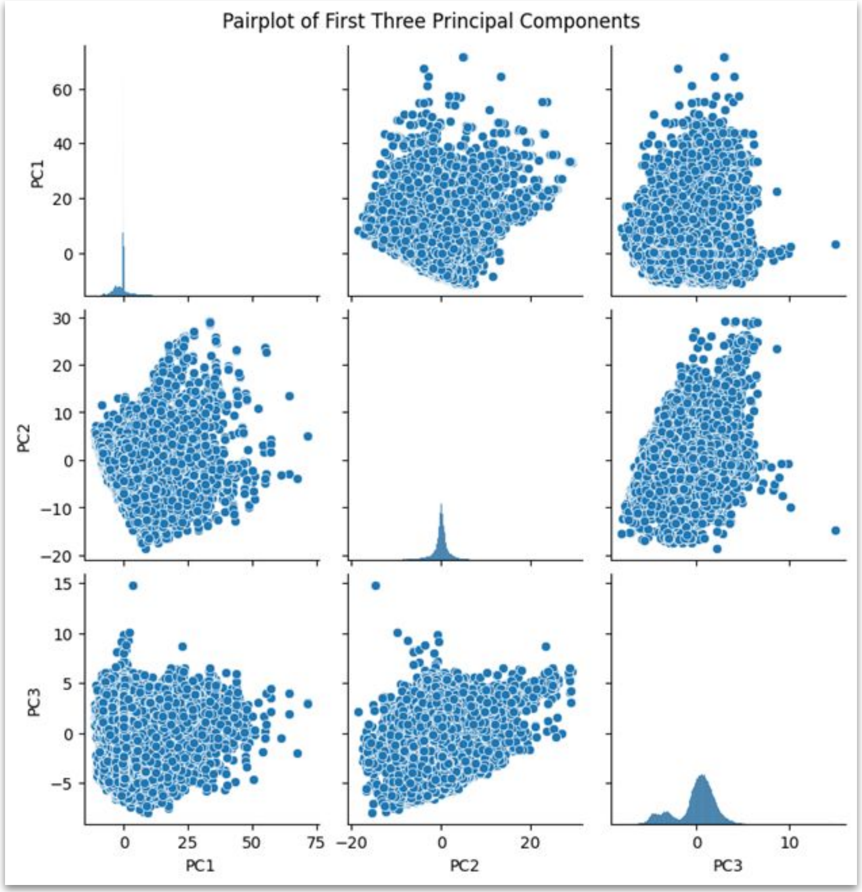
### Dimensionality Reduction

We performed Principal Component Analysis (PCA) on our dataset, aiming to reduce dimensionality while retaining 95% of the variance.

### Model Experimentation

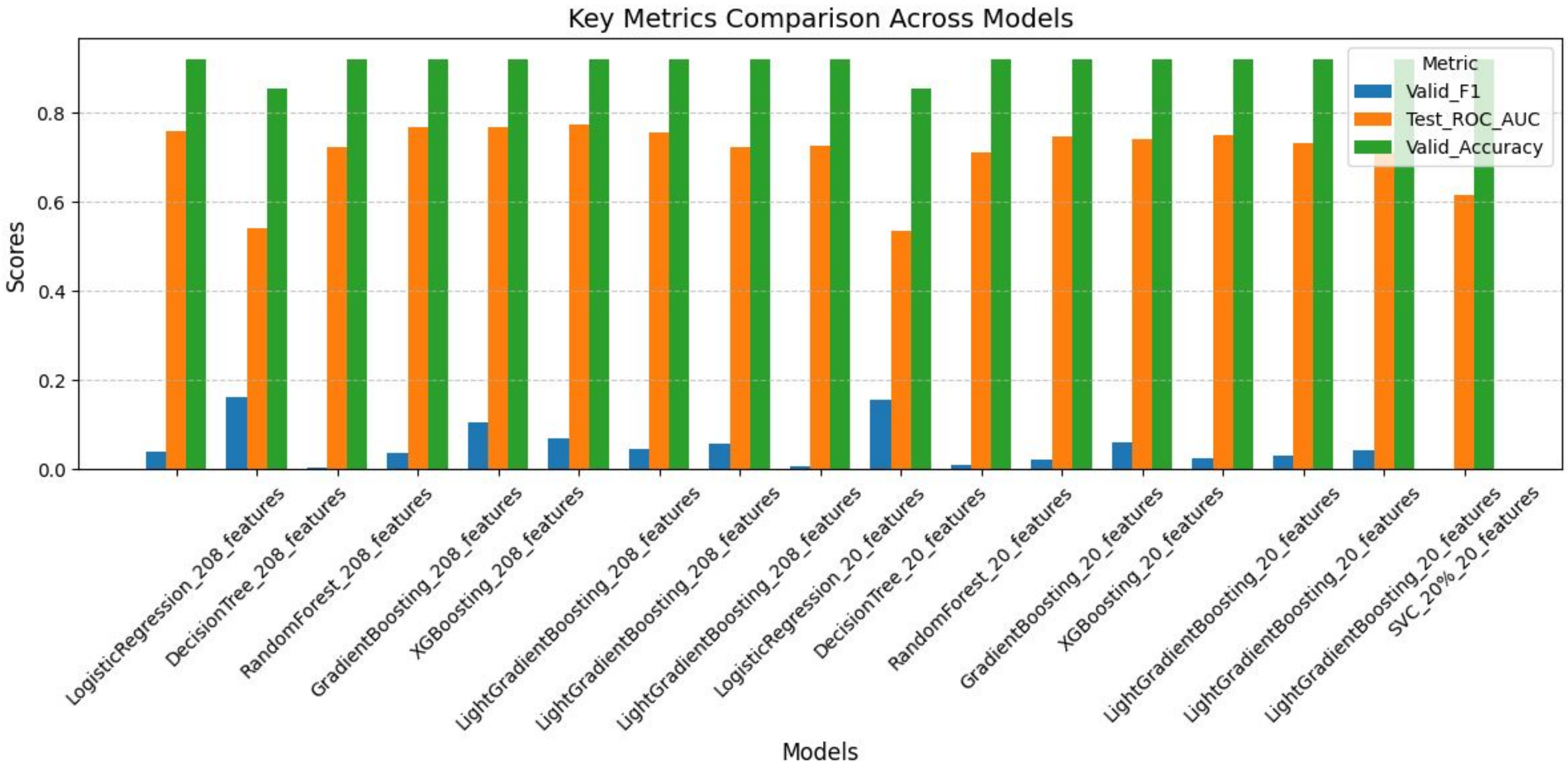
We experimented using 7 different models as well as 2 types of ensemble models on both the full and reduced data sets and then experimented with hyperparameter tuning to find our optimal model.

# Phase 3 Work: Key insights from PCA and experiments




- Performed PCA, but high dimensionality limited its effectiveness.
- Used correlation analysis and feature importance (Decision Trees, Random Forests) to prioritize key features.
- Combined PCA, correlation analysis, and feature ranking into a pipeline, reducing to 20 features with limited performance gains.
- Testing the reduced feature set across models for potential advantages.

# Phase 3 Work: Pipeline results



# Plans for Phase 4

- Develop a more robust modeling pipeline:
  - More types of models, more ensembles
  - Additional hyperparameter tuning
  - Implement and experiment with Neural Networks
- Adjustments or new ideas based on Phase 3 findings
  - Upon completion of Phase 3 we recognize the need for further model types and additional processing power to speed up models training
- Stretch goals
  - We would like to try using all of the data
  - We would like to further refine our metrics

# Challenges Faced

- Complex Dimensionality Reduction process
- Computational limitations of Google Colab even with Pro Account
- Challenges with the imbalance of Target variable
- Challenges with multiple people editing the same file