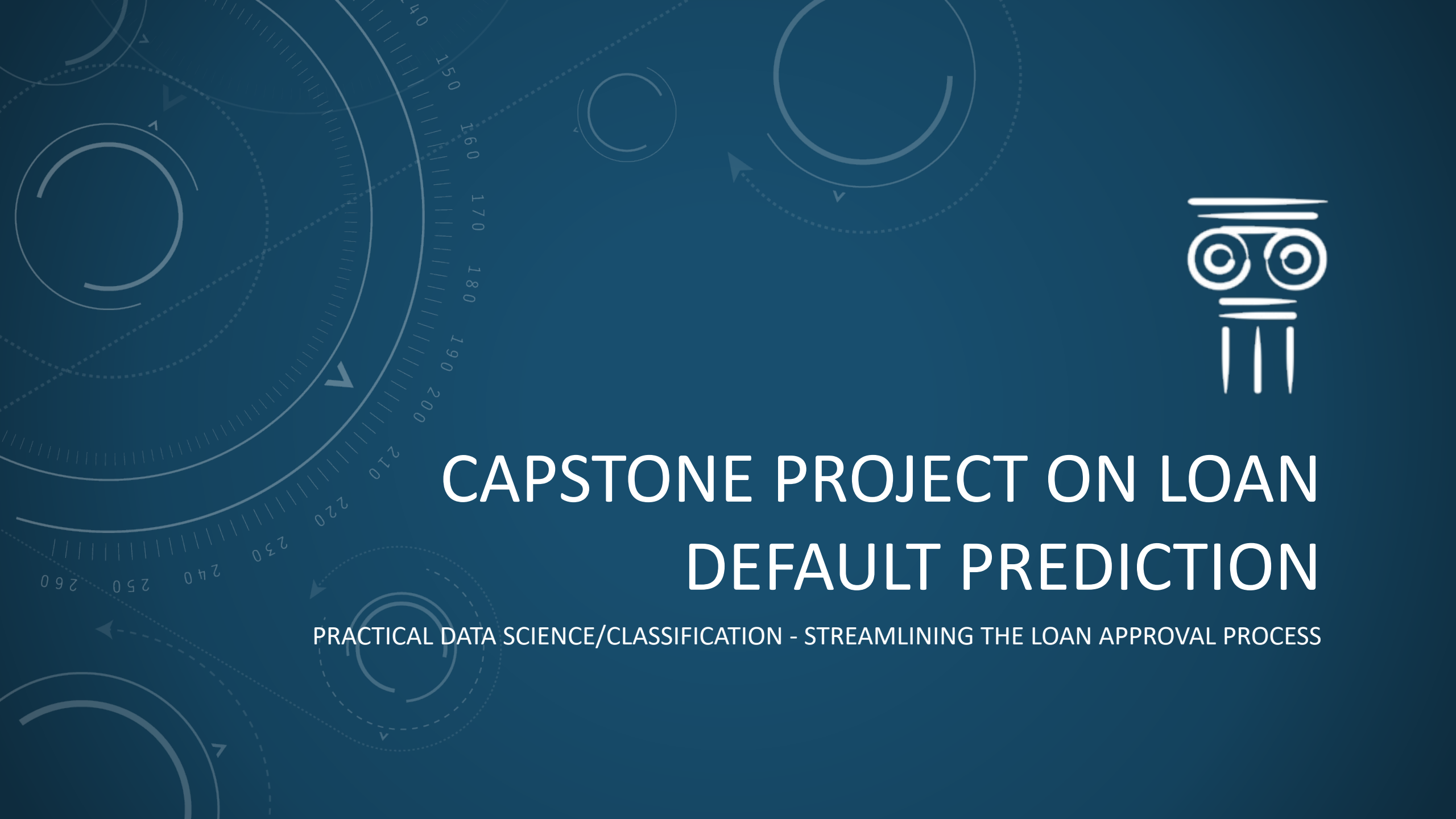
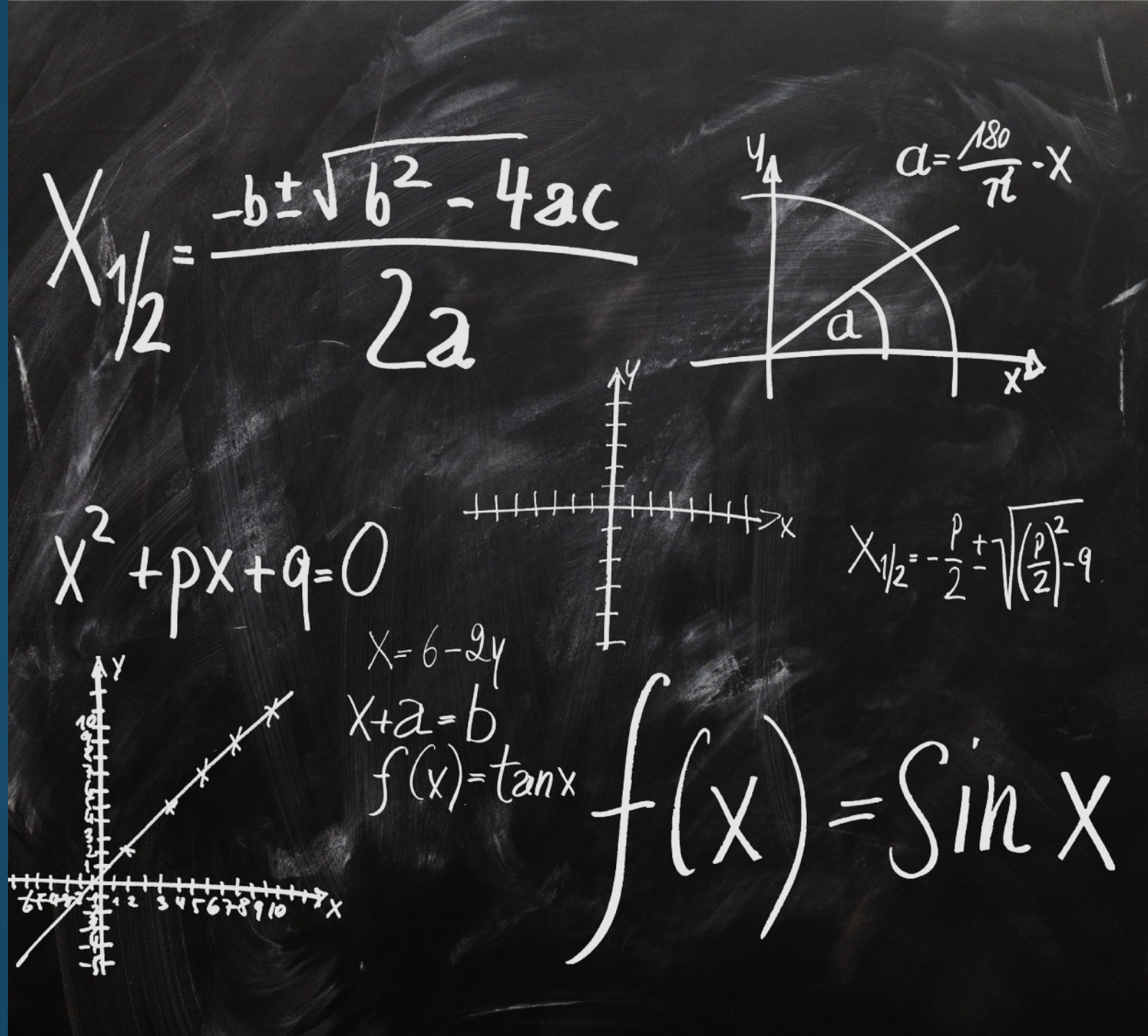


CAPSTONE PROJECT ON LOAN DEFAULT PREDICTION

PRACTICAL DATA SCIENCE/CLASSIFICATION - STREAMLINING THE LOAN APPROVAL PROCESS

AGENDA

1. Problem Definition
2. Data Exploration & Preprocessing
3. Model Selection & Evaluation
4. Key Findings & Insights
5. Proposed Solution & Recommendations
6. Next Steps & Conclusion



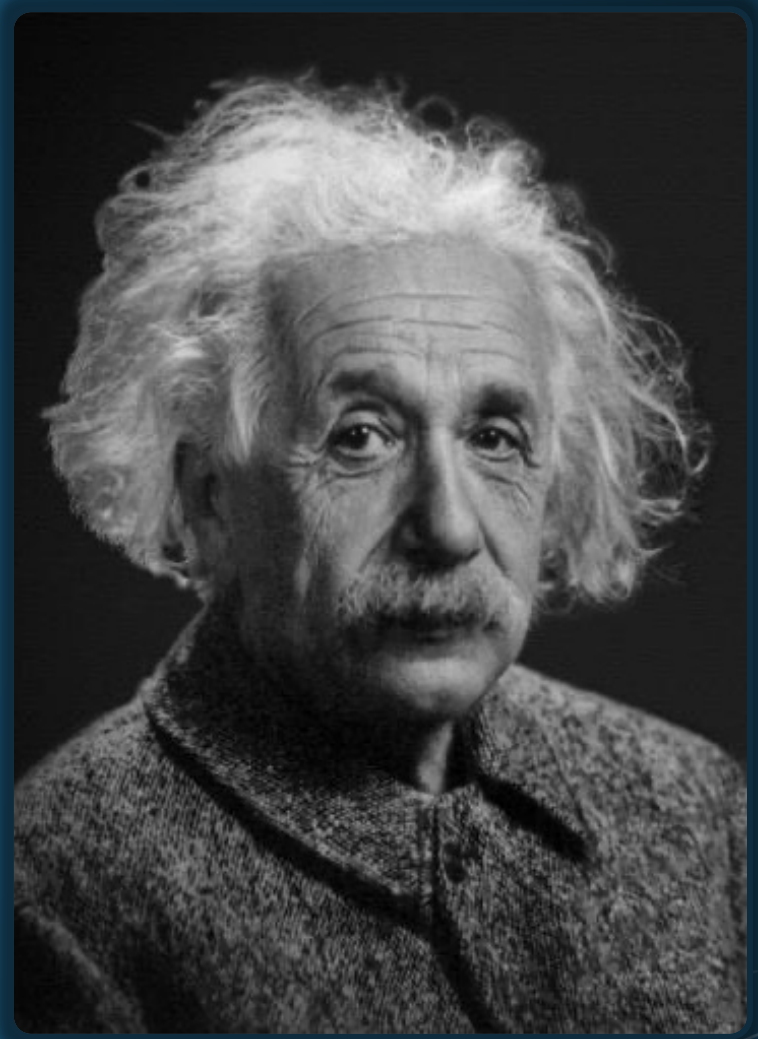
PROBLEM DEFINITION

This project aims to develop a robust credit scoring model for home equity lines of credit, streamlining the approval process by predicting loan defaults and identifying significant features. The model is designed to be unbiased, avoiding past human-centered approval process biases.

- Streamline the decision-making process for approving home equity lines of credit applications
- Develop a robust, accurate, and interpretable credit scoring model
- Leverage data from existing loan underwriting processes
- Predict loan defaults and identify significant features for loan approval
- Ensure the model does not learn biases from past human-centered approval processes

APPROACH FOR SOLUTION

- We developed a predictive model for credit scoring using machine learning, specifically logistic regression and decision tree algorithms.
- The data was preprocessed to improve the model's accuracy, and feature selection was performed to identify the most significant variables.
- We chose machine learning-based approach because it helps in identifying hidden patterns and relationships in the data, thereby improving the model's accuracy.

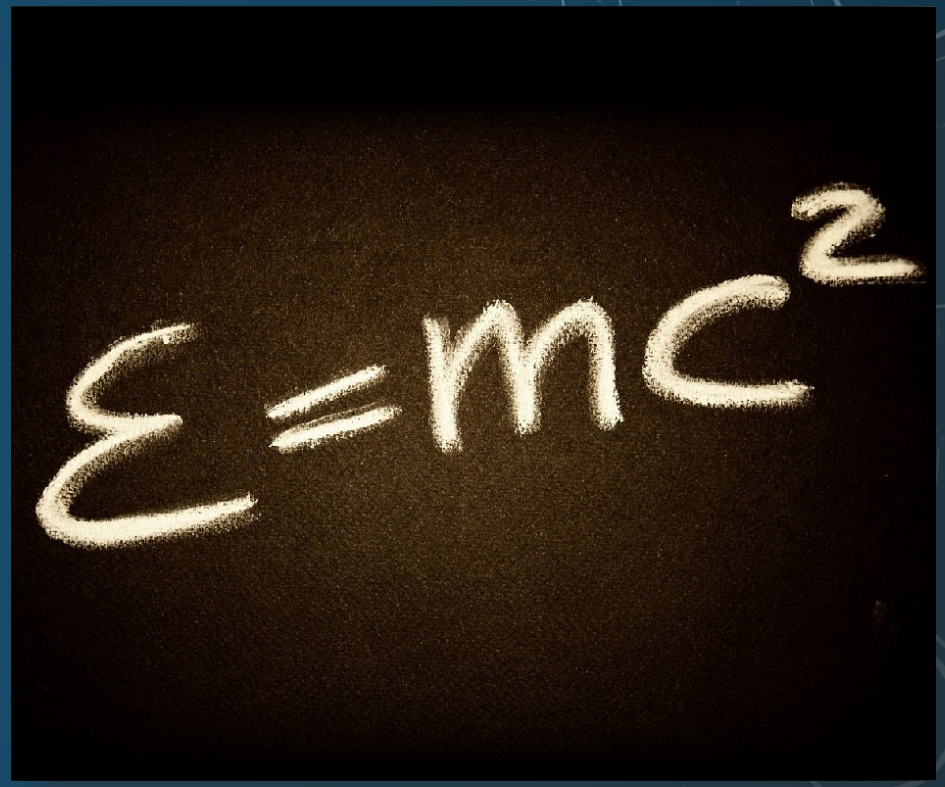


MODEL DEVELOPMENT & EVALUATION

- Models used: logistic regression, decision trees, random forest classifiers
- We evaluated the performance of the models using various measures such as accuracy, recall, precision, and F1-score.
- The random forest classifier model outperformed the other models in terms of accuracy and precision on both the training and test datasets.
- The hyper-tuned decision tree model had better generalization ability, but its accuracy was slightly lower than the random forest classifier model.
- Best performing model: Tuned Random Forest Classifier
  - Training accuracy: 0.996, Test accuracy: 0.907
  - Training recall: 0.981, Test recall: 0.689
  - Training precision: 0.999, Test precision: 0.865

KEY INSIGHTS

- Most significant variables for predicting credit risk:
  - DEBTINC (debt-to-income ratio)
  - CLAGE (loan age)
  - DELINQ (number of delinquent credit lines)
  - CLNO (number of credit lines)
  - YOJ (years at current job)
  - DEROG (number of derogatory reports)
- Missing data flag variables have predictive value
- Logistic regression model insights:
  - DELINQ, NINQ, and DEROG have a significant positive effect on credit risk
  - YOJ has a significant negative effect on credit risk



PROPOSED SOLUTION

- Based on the model's performance and feature importance analysis, we propose adopting the Random Forest Classifier after hyperparameter tuning as the best solution for credit scoring.
- The Random Forest Classifier model is robust to outliers, non-linearity, and multicollinearity, making it a suitable choice for this problem. Its feature importance metrics can provide insights into the most significant factors influencing loan approval, improving the decision-making process.
- Our proposed solution provides a more efficient and unbiased loan approval process, leading to increased profitability and customer satisfaction for the bank.

- Adopt Tuned Random Forest Classifier as the best solution
- Better generalization ability and similar test accuracy scores
- Captures non-linear relationships between predictors and target variable
- Accounts for missing data flag variables

BUSINESS RECOMMENDATIONS

- We recommend implementing an automated loan approval process that relies on the output of the Tuned Random Forest Classifier model to minimize human biases and errors.
- Review and interpret model output to ensure fair treatment of applicants
- Stakeholders should work together to ensure that the model's output is appropriately reviewed and interpreted, especially in cases where the model suggests rejection, to ensure that applicants are being treated fairly.
- The expected benefits of implementing this solution include more efficient and unbiased loan approval processes, leading to increased profitability and customer satisfaction for the bank.

"

The best way to predict your future is to create it.

"

Peter Drucker

Next Steps:

- Monitor and update the model regularly for accuracy and security
- Explore additional hyperparameters or alternative algorithms for further improvements
- Investigate the impact of different variables on loan approval decisions and identify potential sources of bias
- Address associated problems, such as the impact of loan defaults and changing economic conditions on the bank's profitability and risk profile

THANK YOU.