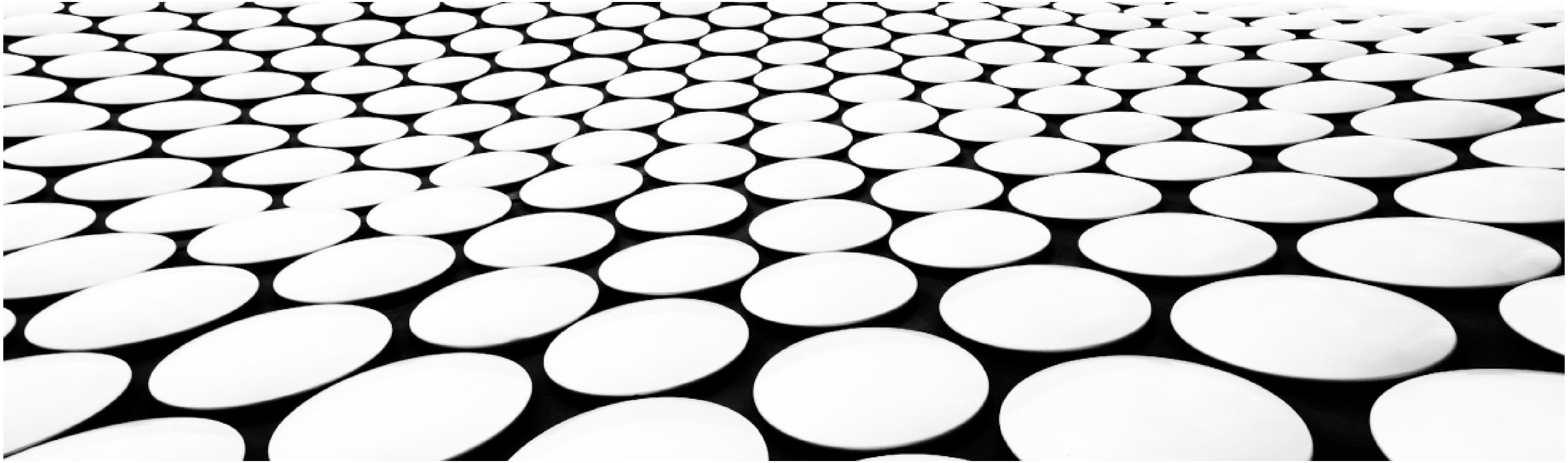
CSCE 5310 FALL 2023

PROJECT NAME:

ANTICIPATE POSSIBLE LOAN DELINQUENCIES AND DETERMINE THE MOST IMPACTFUL DRIVERS



INTRODUCTION OF TEAM, DATA, AND GOALS

Team Members

- Ashwini Kumar Sharma
- Richard Correia
- Panduga Raja Tejasvi Prasad

Dataset

- Comprised of bank data – approximately 67,000 rows x 35 columns of data
- Numerous fields representing loan amounts , term, rates, prior experiences (delinquencies, late fees, etc.), reason for loan (debt consolidation among others) and other values/categories
- Last column is the key – 0 (never became delinquent, 1 (became delinquent
- Distributions are predominantly lognormal

Goals

- Understand the data, distribution qualities and overall connectedness of the data
- Learn what fields are most important and have the biggest impact on deliquency
- Build models using the best fields to predict delinquency

HYPOTHESIS

Null
All of the data parameters (mean and standard deviation), across each column/category, is the same whether the loan status was either a 0 or 1
Implies hard to classify or predict the status of loan

Alternative
One or more of the columns/categories data, supports significant differences between the loan status of 0 and **1**

Fields that were significantly different

- Open accounts, Inquiries within 6 months, Total revolving credit limit, Revolving utilities, and Application type

WORKFLOW

Phase-1

- Pre-process data: detect duplicate and remove null rows
- EDA: descriptive parameters, bar charts/frequency , scatter diagrams (uni and multi), Q-Q plots, correlation checks, etc
- SPSS: analysis for PCA, which columns can be removed
- Assess and restructure data for classification models
- Conclude and summarize
- Write report

Phase-2

- Rerun the existing models with scaling and hyperparameter tuning
- Apply some new models, such as possibly using logistic regression as an example
- Create new fields by possibly taking ratios of the data or further normalizing the data
- Review the analysis and discuss new insights

BREAKDOWN OF WORK COMPLETED: Phase-1

Phase	Description	Action Item	Member	Percentage
1	Research and Exploration	Find Relevant Datasets	Richard	5
		Research papers	Ashwini	5
		Tutorials	Tejasvi	5
2	Thinking exercise for Design part	Finalize Approach	Ashwini	5
		Finalize Hypotheses	Richard	5
3	Implementation - Python	EDA	Ashwini / Tejasvi	20
		Model Training & Validation	Ashwini	10
		Result Comparison	Ashwini	5
	Implementation - SPSS	EDA	Richard	10
		Model	Richard	10
		Hypotheses Results	Richard	5
4	Documentation	Report - First Draft	Tejasvi	15

BREAKDOWN OF WORK COMPLETED: Phase-2

Phase	Description	Action Item	Member	Percentage
3	Implementation	Test accuracy for logistic regression and SVM	Tejasvi / Richard	20
		Hyperparameter tuning	Ashwini	20
		Scaling	Ashwini	20
4	Documentation	Summarization	Richard	30
		Report - Final	Tejasvi	30

CONCLUSIONS – APPLIED MODELS AND DISCRIMINATING CRITERIA

Applied models (with or without over sampling – included because of the unbalanced state of the default status)

- Decision Tree
- Random Forest
- K Neighbors Classifier
- Gaussian NB
- XGBClassifier
- Logistic Regression (phase-2)
- SVM (phase-2)

Performance Criteria

- Highest ROC AUC Score (Area Under the Receiver Operating Characteristic Curve)
- Least Overfitting (Train and Test loss)
- Accuracy

CONCLUSIONS – CHANGES FROM PRELIMINARY TO FINAL

We introduced two new models – SVM and Logistic regression

- both provided less than ideal results compared to other models

Added hyperparameter tuning and scaling

- Tuning improved accuracy
- Scaling provided little impact on results

Based on overall results across all models, XGBoost exhibited the best performance characteristics

RESULTS

Best Performer

XGBoost

With Hyperparameter Tuning on XGBoost

{'max_depth': 7, 'n_estimators': 250}

Results from the Tuning & Scaling

Before tuning
roc_auc_score: 0.837
Train_loss: 3.890
Test_Loss: 5.836}

After tuning
roc_auc_score: 0.970
Train_loss: 0.0794
Test_Loss: 1.062

With scaling
No significant changes

CODE DEMO

End

Thank You