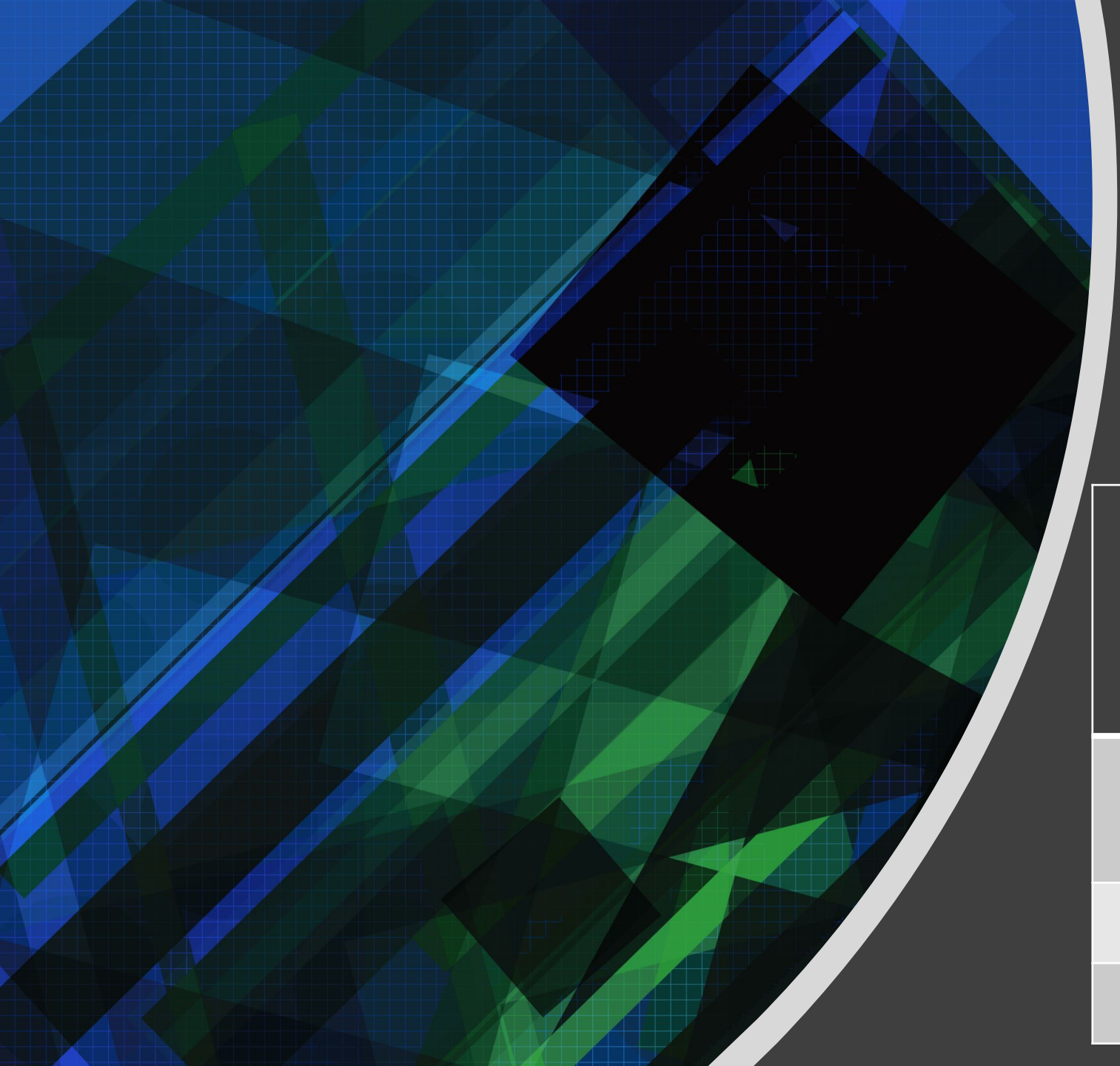
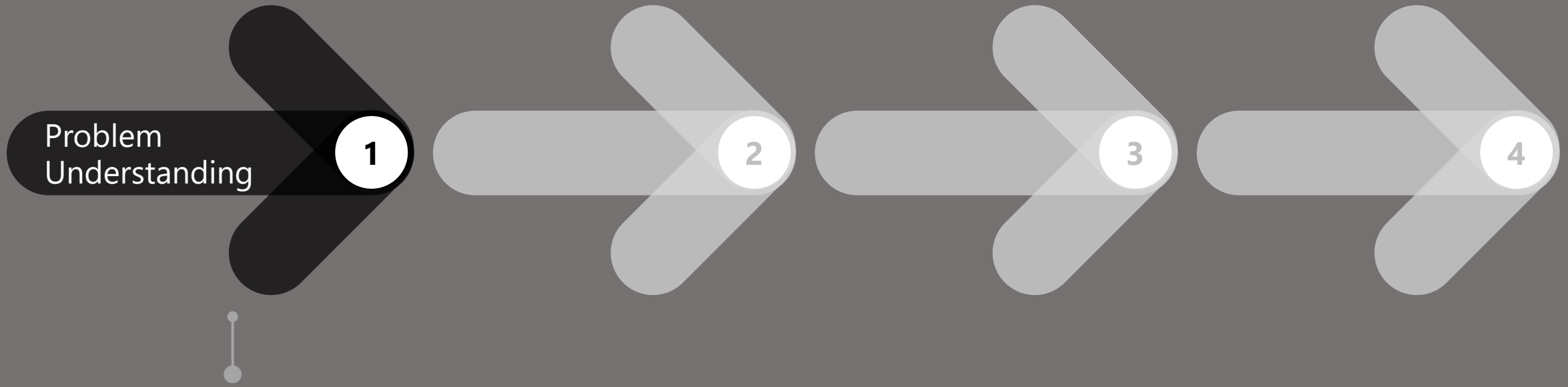


Bankruptcy Prediction

Interim Presentation

Group Members	• Apoorva Garg • Pranavi Krishamshetty • Vishal Pandey • Yash Vahi
Group Number	3
Mentor	Mr. Animesh Tiwari
Team Lead	Yash Vahi

Project Journey : Stage One



In this stage our aim will be to formulate hypothesis, understand the perspective and validate it

Problem Understanding

What is Bankruptcy

- Bankruptcy is a legal process through which people or other entities who cannot repay debts to creditors may seek relief from some or all of their debts. In most jurisdictions, bankruptcy is imposed by a court order, often initiated by the debtor.

Hypothesis Formation and Perspective

- A bank has hired a team of data scientist to predict whether companies taking loans from them are or will go bankrupt in the future and default on their payments to become NPA for the bank.

Broader Utilities of Predicting Bankruptcy

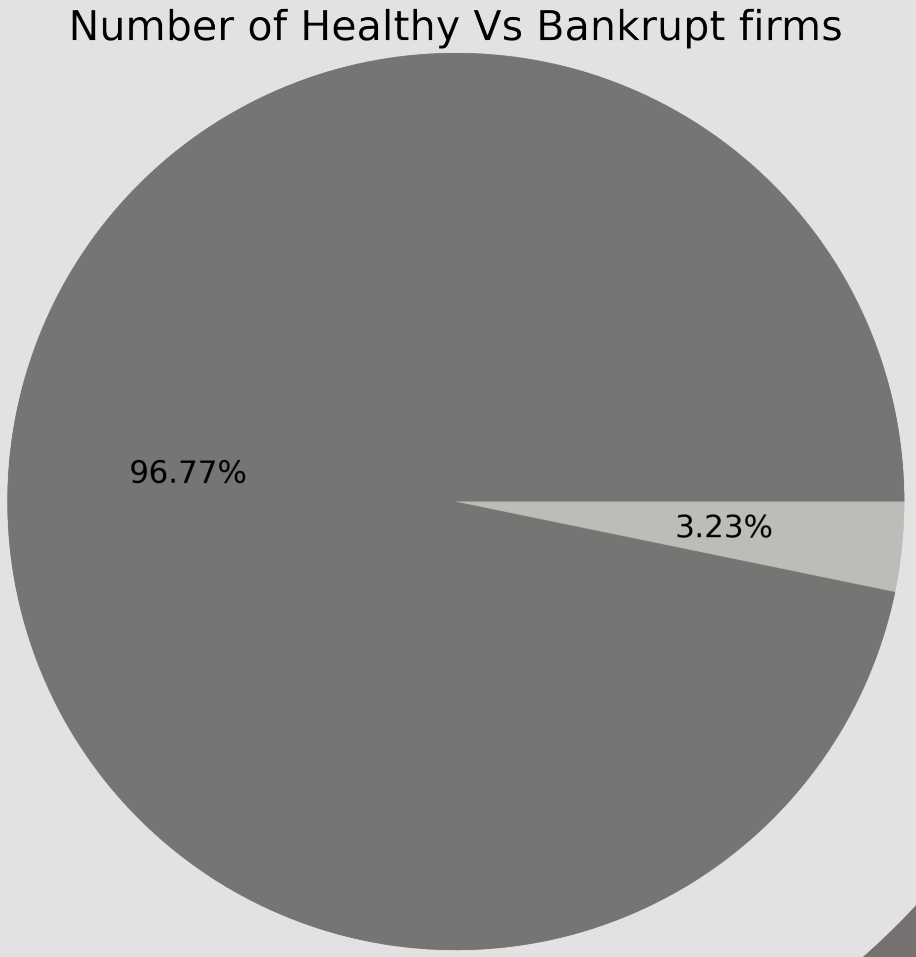
- Predicting bankruptcy is not just beneficial for banks giving loans, companies that are going bankrupt can benefit from identifying tell-tale signs of bankruptcy and stop it in its tracks.

Project Journey : Stage Two

In this stage our aim will be to collect as much information as we can about the data set to make future decisions

Data Description & Basic Analysis

- The Dataset we're working with contains financial information in the form of ratios about companies in Taiwan.
- Shape : 6819 rows and 96 columns
- No null/duplicate values
- All features are numerical
- Columns like "Net Income Flag" and "Liability Assets Flag" are insignificant for further analysis.
- Most features are contained within the range 0-1.



Analysis of Target Variable

Problem

- There are way more Healthy firms than Bankrupt firms in the dataset
  - Almost 97% firms are Healthy
  - Non bankrupt Firms : 6599
  - Bankrupt Firms : 220

Impact

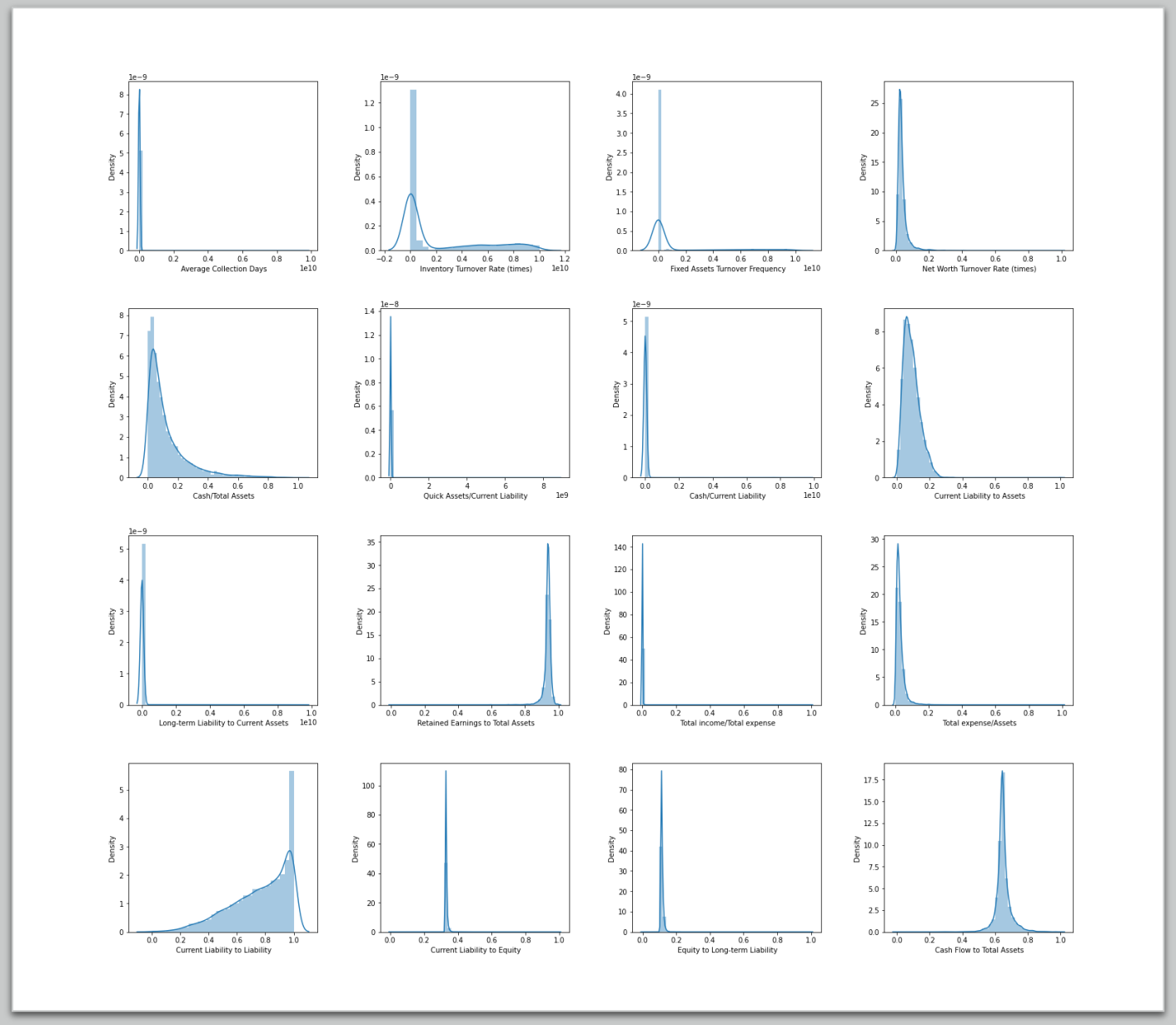
- Class imbalance impacts performance metrics like accuracy ,precision and F1 Score.

Reducing Impact

- Impact can be reduced by under-sampling or over sampling.

Distribution & Outlier Analysis

- Most Features are highly skewed and leptokurtic, pointing towards extreme outliers. Transformation techniques need to be applied
- Our outlier analysis using boxplots did in fact show us that most features have extreme outliers.
- Some common methods of treating outliers are IQR and Zscore elimination



Bivariate and Multivariate Analysis

Correlation Analysis	Bivariate Analysis	Multivariate Analysis
• High-extreme correlation between some of the independent variables. We found 77 distinct pairs of independent features that had correlation greater than 0.7 or less than -0.7.	• Companies with low Research and development expense rate, Tax rate (A),Quick Ratio, Cash/Total Assets, Quick Assets/Current Liability and Cash/Current Liability tend to go bankrupt. • Companies with high Total debt/Total net worth, Inventory Turnover Rate (times),Fixed Assets Turnover Frequency, Allocation rate per person, Current Liability to Assets, Long-term Liability to Current Assets, Quick Asset Turnover Rate, Current Liability to Current Assets and Total assets to GNP price1 tend to go bankrupt.	• Companies with low Net value per share tend to go bankrupt. • Companies with low Persistent EPS in the last four years, Per share net profit before tax, Net profit before tax/paid in capital tend to go bankrupt.

Project Journey : Stage Three

In this stage, we will clean our data and get it ready for modeling.

Data Preprocessing

Outlier Treatment

- We first removed all the outliers using the IQR method.
- We transformed all the features that had absolute skewness > 1 using square root.
- Then we normalized the features to get it ready for KNN imputation.
- At last, we imputed the null values using KNN imputation.

Multicolinearity reduction

- We used the Variance inflation factor(VIF) to filter out features exhibiting multicollinearity(VIF>5).
- We were left with 48 columns.

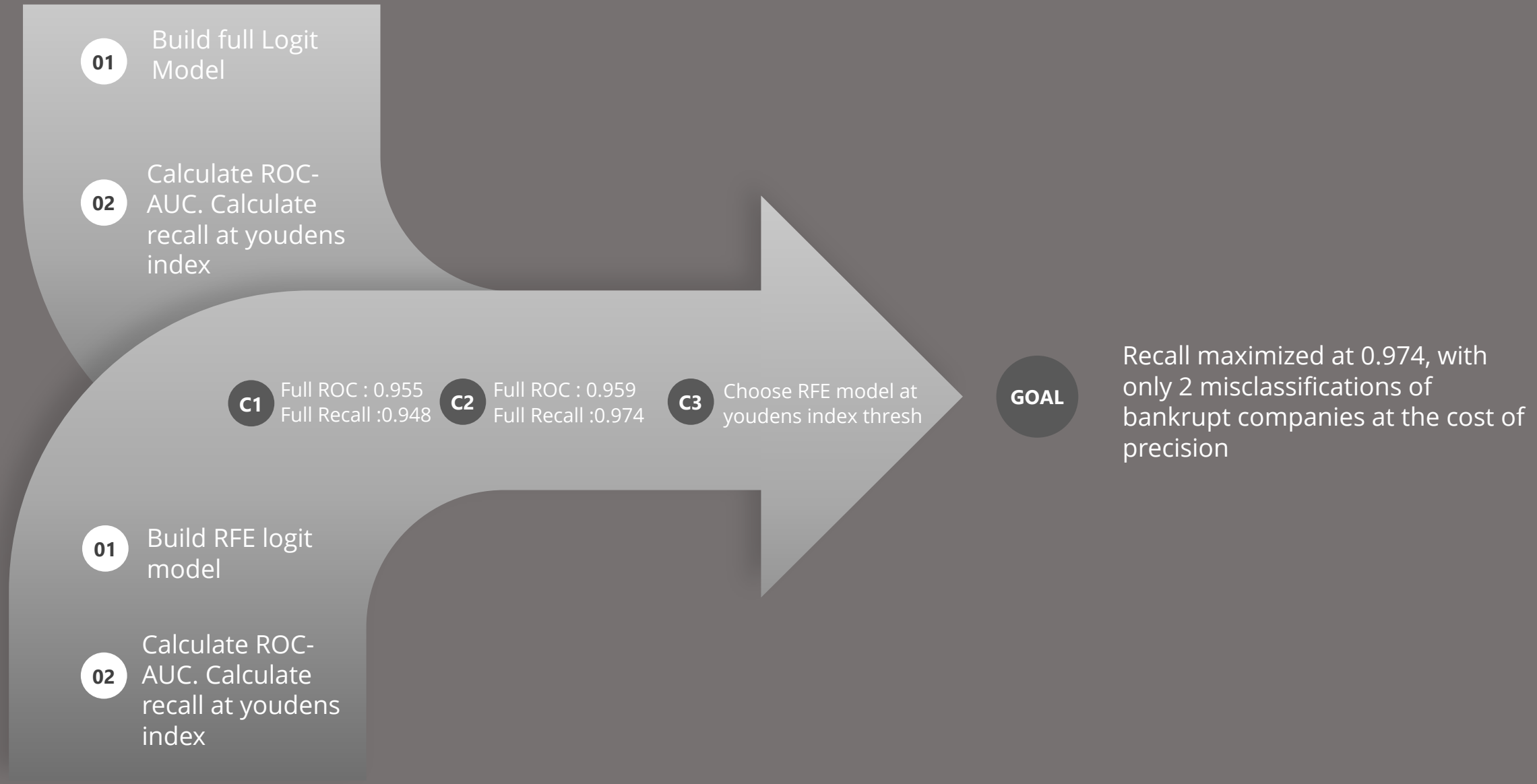
Train-Test split

- We performed a 70:30 split. We also added a constant before splitting.

Project Journey : Stage Four

Now that our Data is ready, we can begin constructing predictive models

Logistic Regression



Recall maximized at 0.974, with only 2 misclassifications of bankrupt companies at the cost of precision

Future Focus

- Build and explore the following models:-
  - Decision Tree
  - Bagging : Random Forest
  - Boosting : Ada, XG, LG (helps improve recall and precision)
  - K-Nearest Neighbors
  - Naïve Bayes
- Compare all our best models using metrics like recall, roc-auc, kappa, f1 with recall being the priority.
- After picking the best model, analyzing coefficients, feature significance/importance.

Thank You!

The floor is open for questions…