SEO UNDERPRICING ANALYSIS

BUSINFO 717: GROUP 1

Dahyun Ryu (dryu783)
Dicky Samudra (dsam041)
Jiajun Li (jli294)
Rao Fu (rfu033)
Regan Ling (rlin870)
Sharan Srinivasan (ssri440)
Xin Zeng (xzen258)

TABLE OF CONTENT

1. Intro & Background

2. Data Preprocessing

3. Machine Learning Model

4. Model Performance

5. Business Insight

6. Conclusion

Background

Underpricing Prediction Model

Objective

- Optimize Pricing Strategies
- Widen Profit Range
- Enhance Market Influences
- Stabilize the Market

SEO (Seasoned Equity Offerings)

Dataset Overview

SEO 2000 - 2009

VARIABLE

Date, Price
Marketplace, Ticker
Issuer, Nation

ISSUE

Missing Values
Duplicate Columns
Format Inconsistency

LIMITATION

Old Dataset
Lack of Important Factors

10443 Rows

112 Columns

SEO Dataset

Data Cleaning

Messy data leads to bad insights.
Clean it, use it, and trust it

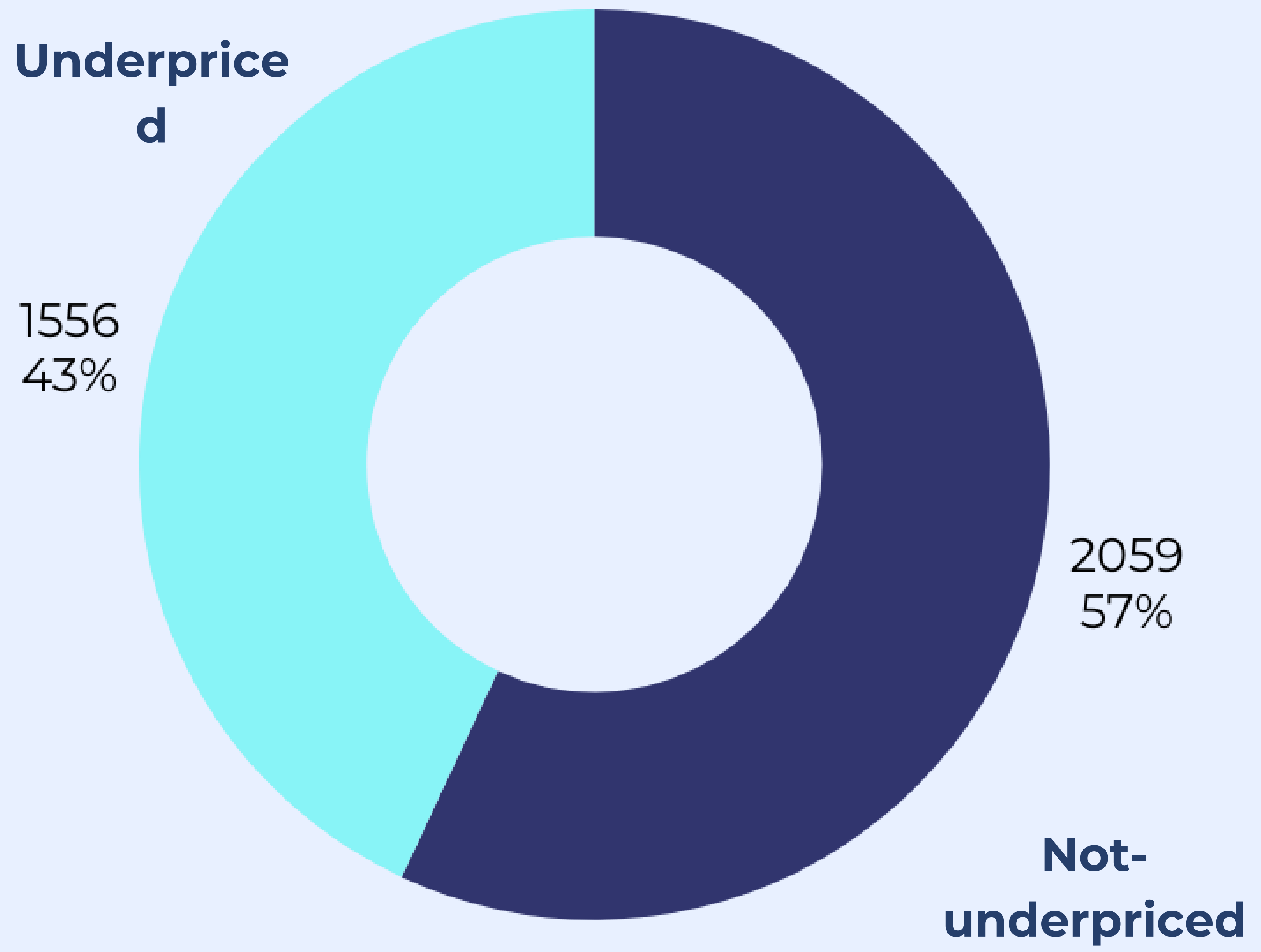
Steps:

- Removes rows and columns with more than 90% missing values
- Drop defines duplicate columns
- Convert date format
- Correct the wrong variable type
- Renaming and Dropping Unnecessary Columns
- Replace NA Values

Underpricing

- **Missing close price data** (retrieved from WRDS)
- **Underpricing calculation** based on offer and close prices
- **2% threshold** used to define underpricing cases
- **Distribution of target variable** (underpriced vs. not underpriced)

$$\text{Underpricing} = \frac{\text{Close Price} - \text{Offer Price}}{\text{Offer Price}}$$



Data Sourcing

Price to Sales Ratio

A measure of a company's sales relative to its market capitalization.

Accrual to Sales Ratio

Reflects the relationship between accruals and sales revenue.

Common Shares Traded

The total number of shares that are traded in a given period.

Market Volatility

Measures the degree of variation in trading prices over time.

Return on Equity

Indicates how effectively management is using a company's assets to create profits.

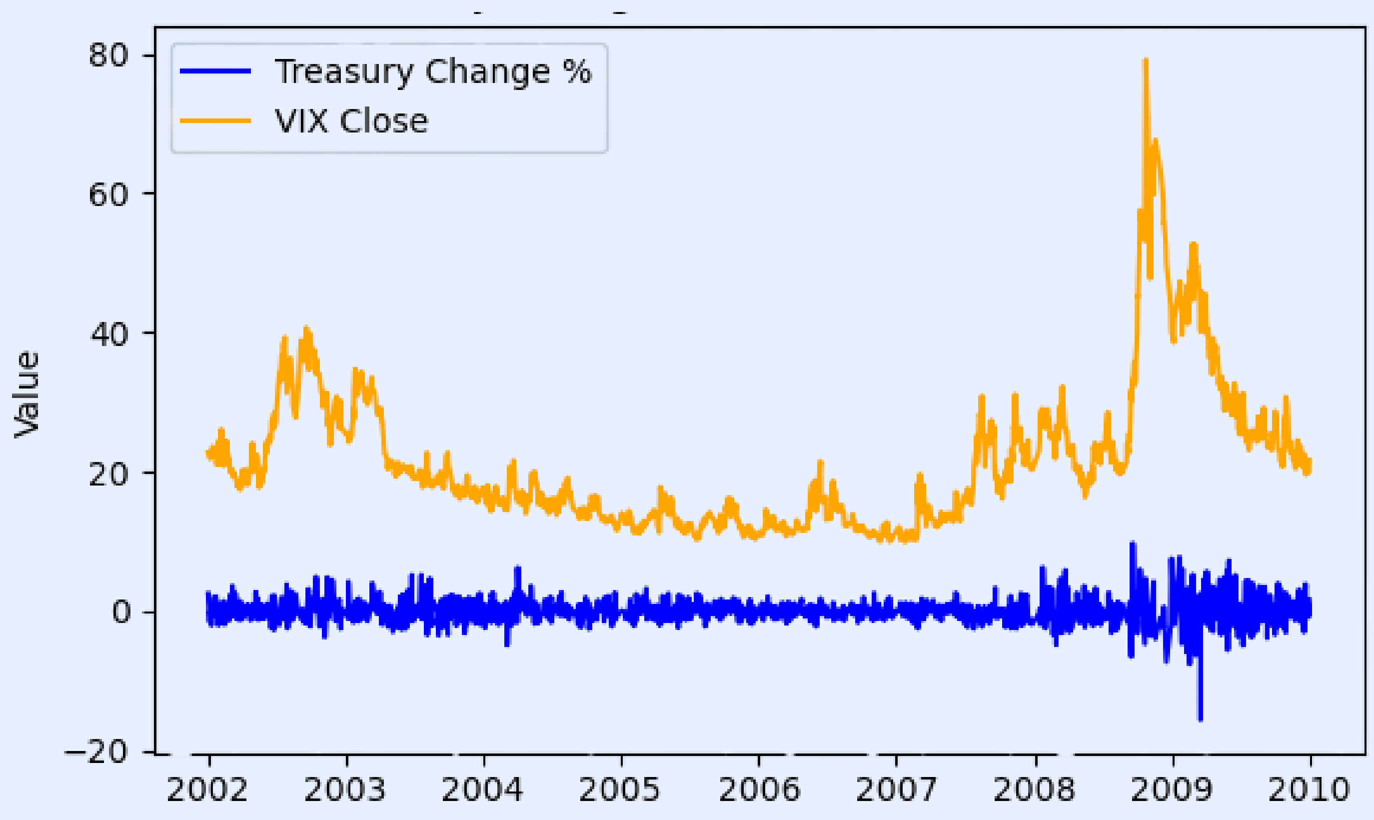
Price to Book Ratio

Compares a company's market value to its book value.

Historical Treasury Yield

Shows the yield on U.S. Treasury securities over time.

Dataset Overview: 3,615 records with **116 variables.**



Data Souce: Wharton Data Reasearch Centre (WRDS), Yahoo Finance and CBOE

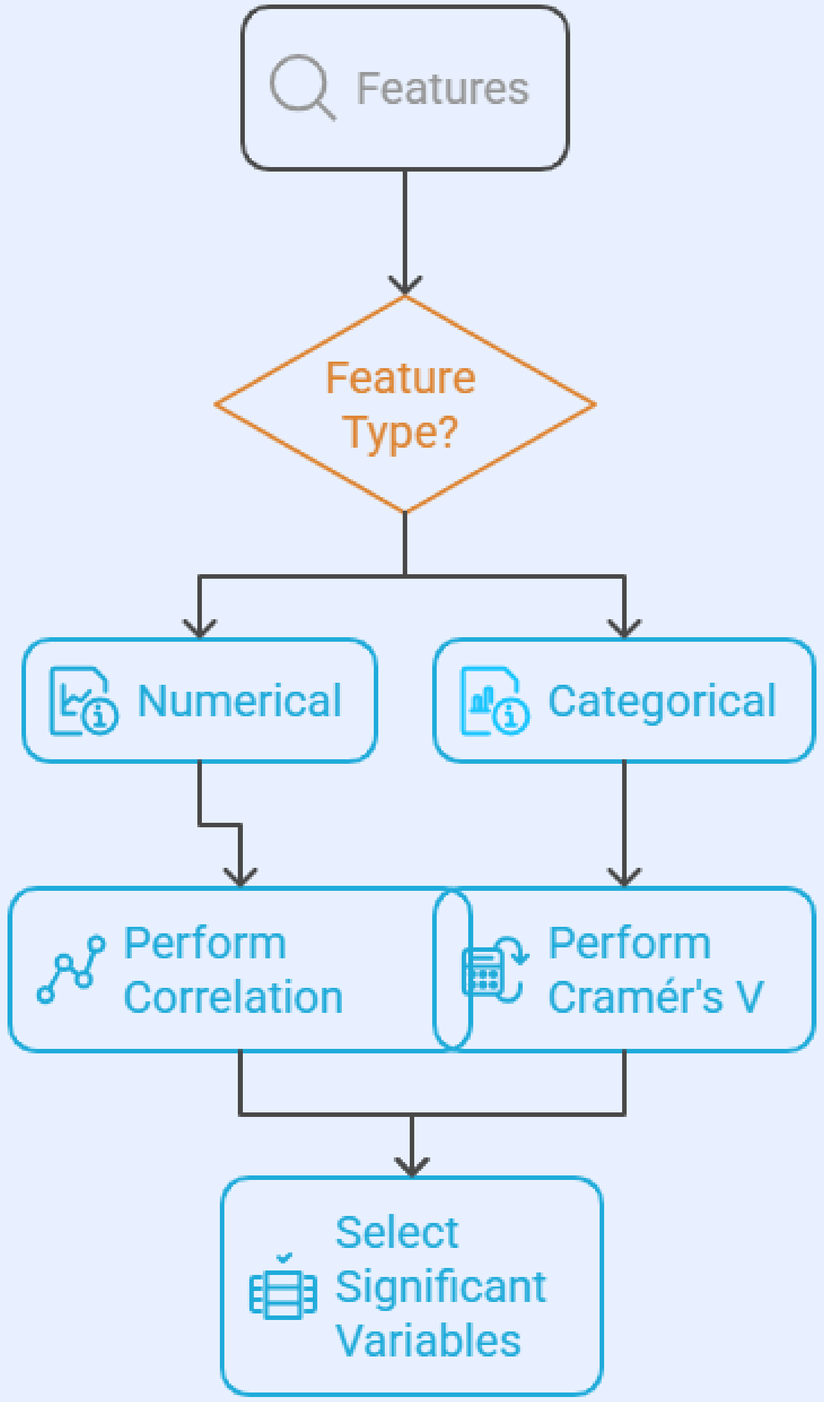
Feature Selection

Final Dataset: 29 Features, 3615 observations

Key Features Selected:

- **Deal Type** – Category of the financial deal analyzed.
- **Type of Security** – Specific financial instrument involved.
- **Primary Exchange** – Platform where the security is traded.
- **Shares Filed –** Number of shares registered for trading.
- **Price Close –** Final price at which the security is offered.
- **VIX** – Market volatility index indicating risk.

Feature Selection Process



Model Comparison

Random Forest

✔ Strong for handling complex patterns

❗ Can be slow for big data

👉 Best for structured data with non-linear patterns

LightGBM

Fast and efficient for large datasets

Sensitive to tuning parameters

Best for large-scale predictions with speed

Logistic Regression

Simple and easy to interpret

Struggles with complex relationships

Best for clear, linear trends

Key Takeaway:
LightGBM and Random Forest perform best for **data-driven pricing predictions,** while Logistic Regression offers **interpretability but limited flexibility.**

Understanding Classification Decisions

How Do We Measure Performance?

Thresholds in Classification

- **Higher threshold →** Fewer false positives, may miss real cases.
- **Lower threshold →** Catches more positives, increases false alarms.
  - Example: Lower thresholds in fraud detection may block real transactions.

Cross-Validation (10 Folds)

- Splits data into **10 parts**, trains on 9, tests on 1.

Why does this matter?

- Helps businesses **reduce financial risks** and **make smarter predictions.**

Model Performance Comparison

Model	Accuracy	Recall	Precision	True Positive	False Positive
Random Forest	**63.76%**	**77.17%**	**55.68%**	**240**	**191**
LightGBM	61.85%	55.30%	55.66%	172	137
Logistic Regression	60.72%	61.41%	53.72%	191	164

Confusion Matrix (Random Forest)

Actual Label \ Predicted Label	Negative	Positive
Negative	221	191
Positive	71	240

Why are these important?

Accuracy: 63.79%

Trustworthiness & Consistency

TP > FP (49 cases differences)

Actionable Insights

Precision: 55.68%

Resource Efficiency

Recall: 77.17%

Risk Identification

MONETISATION STRATEGY & RISK IDENTIFICATION

Plan to generate profits using the model

1. **Provide the model with data from a company currently engaged in SEO to assess whether it is undervalued.**

2. **Allocate capital to purchase shares at the offer price.**

3. **Low cut off value: we target small margin but we play in huge volume.**

4. **Sell these shares on Day 1 after the offering, once the market has "priced in" the information.**

Risk Implementing ML model

- **Limited accuracy** may lead to financial losses. **However, risk & reward remain positive.**
- **False negatives** could miss undervalued assets (~30%).
- **Lack of generalization**—model may struggle with future data.
- **Liquidity risk** due to high trading volume.

LIMITATIONS

01 Noise in Financial Data

02 Lack of Explainability in ML Models

03 Limited Timeframe Evaluation

FUTURE IMPROVEMENTS

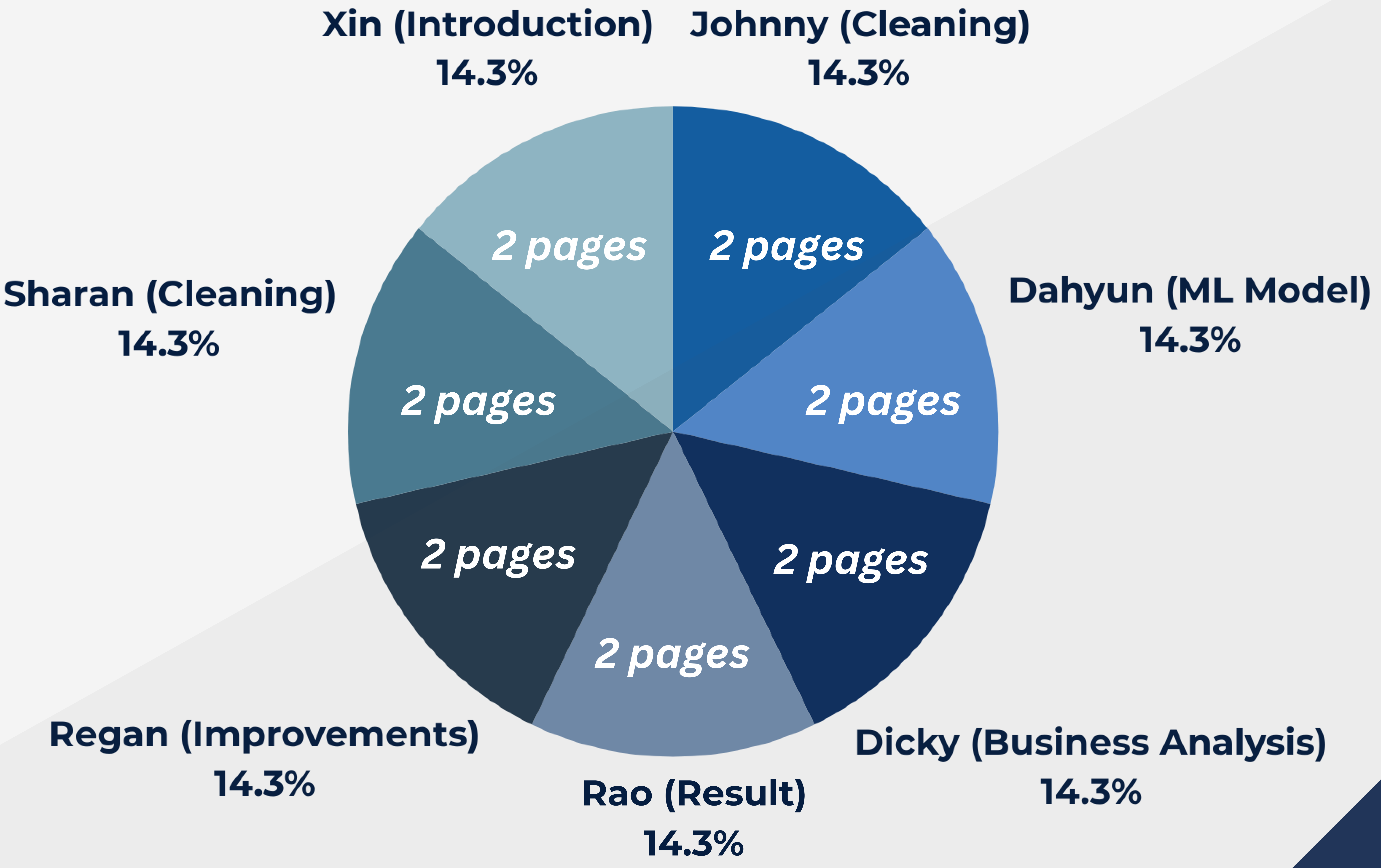
BackTesting

Backtest the model
for longer timeframe

Improving Model Explainability

- SHAP (Shapley Additive Explanations)
- LIME (Local Interpretable Model-Agnostic Explanations)
- Using simpler, interpretable models in parallel

Team Contribution



Team Contribution

Member	Description	Slide Pages
Xin	Introduced the project, explaining the objective of optimizing pricing strategies, improving underwriting ability, and enhancing market stability.	3-4
Sharan	Worked on data preprocessing, including cleaning the dataset by handling missing values, removing duplicate columns, and ensuring consistency in data formatting.	5-6
Jiajun	Focused on further data cleaning, ensuring the dataset was structured correctly for model training, and preparing it for feature selection.	7-8
Dahyun	Developed and implemented machine learning models, comparing different algorithms such as Random Forest, LightGBM, and Logistic Regression.	9-10
Rao	Analyzed the model performance, comparing accuracy, recall, and precision, and presented the results using confusion matrices and key performance metrics.	11-12
Dicky	Conducted the business analysis, interpreting the model's practical applications, potential risks, and the strategy for monetizing insights from SEO underpricing predictions.	13-14
Regan	Addressed limitations and proposed future improvements, such as backtesting over a longer timeframe and refining feature selection to enhance model efficiency.	15-16

References

Altınkılıç, O. and Hansen, R.S. (2003). Discounting and underpricing in seasoned equity offers. *Journal of Financial Economics, 69*(2), 285-323. https://doi.org/10.1016/s0304-405x(03)00114-4.

Corwin, S. A. (2003). The determinants of underpricing for seasoned equity offers. *The Journal of Finance, 58*(5), 2249–2279. https://fbe.unimelb.edu.au/__data/assets/pdf_file/0008/3802229/Working-Paper-Series_Letter_James-Brugler_14_20.pdf

Kim, W., & Shin, H. H. (2004). The puzzling increase in the underpricing of seasoned equity offerings. *Financial Review, 39*(3), 343–365. https://colab.ws/articles/10.1111%2Fj.0732-8516.2004.00079.x

THANK YOU

We are welcome any question