Financial market trend recovery

Manaswini Pedimalla

November 04, 2024

1 Introduction

This project aims to predict regulatory changes in the financial market, helping compliance teams anticipate policy shifts. Using a Kaggle dataset with daily financial metrics like Market Price, Volatility Index, and Recovery Index, we identify patterns associated with regulatory adjustments. By building and evaluating classification models, this tool provides early insights to support strategic planning and enhance regulatory preparedness.

2 Stakeholders

Our main stakeholder for this project is the financial analysis team at an investment firm. Their focus is on understanding and forecasting market recovery trends to better manage risks and make informed investment decisions during periods of economic volatility.

3 Problem Statement

The goal of this project is to create a machine learning model that predicts the likelihood of market recovery on any given day, based on historical data. By having a reliable prediction, the financial analysis team can better understand recovery patterns, helping them anticipate market trends and make informed choices during uncertain times.

4 Dataset

This dataset, which simulates financial market trends and recovery patterns, is available on Kaggle under the category of financial market simulation data. It was designed to mimic daily market activities, particularly focusing on recovery and volatility after economic downturns.

5 Models Tested

I tested four models:

- **Logistic Regression**: I chose this model as a baseline due to its simplicity and interpretability, which makes it easier for stakeholders to understand the influence of each variable on predictions. Logistic Regression is effective for binary classification tasks, providing initial insights into the problem.

- **Random Forest Classifier**: This model is a powerful ensemble method that captures complex interactions between features. I chose Random Forest because it handles diverse data well and offers high accuracy, making it valuable for a potentially high-impact application like regulatory forecasting.
- **Naive Bayes**: I included Naive Bayes due to its efficiency and robustness with smaller datasets, even under the assumption of feature independence. This model provides probabilistic outputs and often serves as a quick yet effective option for classification.
- **Decision Tree Classifier**: The Decision Tree model was selected for its transparency. The decision rules generated by the model allow the stakeholder to understand the logic behind predictions, which is particularly useful for interpretability in a regulated industry.

These models were chosen to balance complexity, interpretability, and accuracy, providing insights into both baseline performance and advanced capabilities.

6 Feature Selection/Engineering

To enhance the model's predictive accuracy, I engineered the following features:

- **Profit Margin** (calculated as Profit ($) / Revenue ($)): This feature provides insights into profitability relative to revenue, potentially uncovering trends in financial health that could signal impending regulatory changes.
- **Expense-to-Revenue Ratio** (calculated as Expense ($) / Revenue ($)): This metric evaluates operational efficiency, which may correlate with areas of compliance risk or highlight vulnerabilities in the face of regulatory shifts.

In addition to these new features, I selected core financial metrics such as Revenue ($), Profit ($), Market Price ($), and Volatility Index. These metrics are fundamental indicators of market stability and are likely factors that regulatory bodies monitor. The engineered and selected features were chosen based on their financial significance and their potential to offer stakeholders valuable insights into market conditions and regulatory readiness.

7 Hyperparameter Tuning

We used **Grid Search** to systematically test combinations of these parameters, selecting the best settings based on model performance on the validation set.

1. **Logistic Regression**: Tuned the regularization strength (C) and solver types (liblinear, saga, newton-cg) to optimize the model's performance and computational efficiency.

2. **Random Forest Classifier**: Adjusted the number of trees (n_estimators), tree depth (max_depth), and minimum samples for splits (min_samples_split) to improve accuracy and prevent overfitting.

3. **Decision Tree Classifier**: Tuned the criterion (gini, entropy), tree depth (max_depth), and minimum samples per leaf (min_samples_leaf) to avoid overfitting and enhance model performance.

4. **Naive Bayes**: Focused on adjusting prior probabilities, though it is generally less sensitive to hyperparameter tuning compared to other models.

8 Model Evaluation

Precision and recall are particularly important in financial contexts, where predicting recovery incorrectly could lead to risky decisions. The F1 score provides a single measure that captures the balance between these two metrics.

Evaluation Metrics:

1. **Precision**: Precision was important to reduce false positives (incorrectly predicting recovery), as a false positive might imply stability when the market remains volatile.

2. **Recall**: We also prioritized recall to ensure we catch as many actual recovery days as possible, minimizing the chance of missing a true recovery.

3. **F1 Score**: We used F1 Score as a balanced metric, combining precision and recall, which helps us get an overall sense of model performance for the binary classification task.

4**. Accuracy**: This provided an overall measure of the model's correctness, useful as a starting point to understand the general performance across all models.

9 Results and Analysis

Upon evaluating the performance of each model, the following results were observed:

The Random Forest Classifier achieved the best performance with 78% accuracy and an F1 score of 0.78, making it ideal for market recovery prediction. Logistic Regression had moderate performance, while Naive Bayes underperformed. Hyperparameter tuning improved the Decision Tree, but it still lagged behind Random Forest. I recommend using Random Forest for its strong precision and recall, with potential for further improvement through optimization and feature engineering.

10 Future Work

For future improvements, adding features like moving averages and using advanced models such as Gradient Boosting or XGBoost could enhance predictions. Tuning the models further and integrating real-time data would help optimize performance. Deploying the model for real-time predictions would enable better decision-making in changing market conditions.

11 Client Recommendation

I recommend the client use the Random Forest Classifier for market recovery prediction, as it provided the best results in accuracy, precision, and recall. Further tuning and feature engineering can enhance its performance. Furthermore, integrating real-time data and deploying the model would help make timely, data-driven decisions in dynamic market conditions.

12 Conclusion

In conclusion, the Random Forest Classifier was the most effective model for predicting market recovery, achieving strong performance in accuracy, precision, and recall. Feature engineering and data balancing improved the model's robustness. While further optimization and real-time data integration could enhance performance, the model provides valuable insights. This can help stakeholders make informed decisions and navigate market changes more effectively.

13 Dataset and Github Links

1. Dataset: https://www.kaggle.com/datasets/ziya07/financial-market-trend-recovery
2. Github: https://github.com/mpedimal2816/Ml-Project-1