Feature Optimization for Logistic Regression Models for Netflix Stock Price Prediction

Shourya Mukherjee
Department of Computer Science
University of North Carolina at Chapel Hill
smukherjee@unc.edu

Dhiren Gangishetty
Department of Computer Science
University of North Carolina at Chapel Hill

Snehashish Reddy Manda
Department of Computer Science
University of North Carolina at Chapel Hill

Girish Rengadurai
Department of Computer Science
University of North Carolina at Chapel Hill

Dhyey Shah
Department of Computer Science
University of North Carolina at Chapel Hill

Abstract

This study investigates the predictive power of logistic regression models for forecasting Netflix stock price movements using both traditional financial indicators and content-related features. By incorporating data on the volume of high-rated titles added to Netflix's library, we extend beyond conventional stock prediction approaches. Our optimized logistic regression model achieves 96.4% accuracy on test data, with balanced performance across both upward and downward price movements. Cross-validation across 500 iterations confirms the model's robustness, yielding a mean accuracy of 90.55%. The strong correlation between high-rated content additions and stock performance highlights the value of integrating content quality metrics into financial prediction models. These findings demonstrate that relatively simple machine learning approaches, when paired with domain-specific features, can effectively predict stock price movements in the streaming entertainment sector.

Introduction

Predicting stock price movements has been a focal point of financial research for decades, with machine learning approaches gaining significant traction in recent years Henrique et al. [2019]. The stock market's inherent complexity and volatility make it a challenging yet rewarding domain for predictive analytics Atsalakis and Valavanis [2009]. Among the various market sectors, the entertainment streaming industry, particularly Netflix, presents a unique case study due to its rapid growth, content-driven business model, and sensitivity to both market trends and competitive pressures Gomez-Uribe and Hunt [2016].

Netflix, as one of the pioneering streaming services, has transformed from a DVD rental business to a global entertainment powerhouse with over 200 million subscribers worldwide Netflix, Inc. [2023]. This evolution has been accompanied by significant stock price fluctuations, making it an interesting target for predictive modeling. Traditional financial analysis relies heavily on technical indicators derived from historical price and volume data Murphy [1999], while fundamental analysis considers company-specific information and broader economic factors Graham and Dodd [2006].

Our research extends beyond conventional financial metrics by incorporating content-related features, specifically the addition of high-rated titles to Netflix's library. This approach is motivated by research suggesting that content quality and library expansion significantly impact subscriber retention and acquisition, which in turn influence investor sentiment and stock performance Hassouna et al. [2020]. By combining technical indicators with content-related features, we aim to capture a more comprehensive view of the factors driving Netflix's stock price movements.

Recent advancements in machine learning have demonstrated promising results in stock price prediction tasks Patel et al. [2015]. Logistic regression, despite its simplicity compared to more complex models, has shown competitive performance in binary classification tasks related to financial markets Ballings et al. [2015]. Its interpretability makes it particularly valuable for understanding the relative importance of different features in the prediction process James et al. [2013].

In this study, we employ logistic regression with careful hyperparameter tuning to predict the monthly directional movement of Netflix stock prices. We evaluate our model using standard classification metrics and perform principal component analysis to identify the most influential features. Our approach builds on previous work by Li et al. [2016] on stock trend prediction and West [2000] on credit scoring using logistic regression, adapting these methodologies to the specific context of Netflix stock prediction with the addition of content-related variables.

Methods

Dataset

We use a dataset containing features related to Netflix stock prices and the volume of high-rated titles added over time. The target variable is the *Monthly Price Movement* of Netflix stock, where the task is to predict whether the stock price will go up or down in the subsequent month based on the features of the current month. The dataset includes the following key features: the opening price (*Open*), closing price (*Close*), the previous month's closing price (*Prev_Close*), the high price of the stock (*High*), the low price of the stock (*Low*), the volume of stock traded (*Volume*), and the number of high-rated titles added in the previous month (*High_Rated_Titles_Added_Last_Month*).

Data Preprocessing

The raw dataset was preprocessed as follows:

- **Feature Selection**: A subset of features was selected for the predictive model, including stock-related features (*Open*, *Close*, *Prev_Close*, *High*, *Low*, *Volume*) and a time-series feature related to the volume of highly rated Netflix titles added in the previous month (*High_Rated_Titles_Added_Last_Month*).
- **Data Splitting**: The dataset was split into training and testing sets using an *70/30 split*. Specifically, 70% of the data was used for training, while 30% was reserved for testing. This ensured that the model is validated on unseen data, preventing overfitting.
- **Feature Scaling**: Feature scaling was performed using the `StandardScaler` from scikit-learn. This scaler standardizes the features by removing the mean and scaling to unit variance. The scaler was fit on the training data and subsequently applied to both the training and test datasets to ensure that no information from the test set leaked into the training process.

Model Selection and Hyperparameter Tuning

We employed *Logistic Regression* as the primary model for binary classification of the stock price movement. Logistic Regression is a widely used linear classifier suitable for problems like this, where the output is a binary variable. The model predicts the log-odds of the outcome, making it appropriate for predicting binary events like price movement (up or down).

The **hyperparameters** of the Logistic Regression model were optimized using `GridSearchCV` with **cross-validation**. Specifically, we performed grid search over the regularization parameter C (ranging from 1×10^{-4} to 1×10^{4}) and selected between two solvers, `lbfgs` and `liblinear`. The hyperparameter search was done using *5-fold cross-validation*, ensuring that the model's performance was robust and generalized across different data splits. `GridSearchCV` also ensures that the best combination of hyperparameters is selected based on cross-validation accuracy.

Model Training

After hyperparameter tuning, the model with the best parameters was refitted on the entire training dataset. The final model was trained using the best combination of regularization strength (C) and solver determined by `GridSearchCV`. This ensured that the model was trained in the most optimal configuration.

Model Evaluation

Model performance was evaluated on the *test dataset* that was held out during training. The following metrics were used to assess the model's accuracy and predictive performance:

- **Accuracy**: The overall accuracy of the model on the test set, which represents the proportion of correct predictions.
- **Classification Report**: The classification report provides detailed metrics, including precision, recall, F1-score, and support for each class (Up/Down). These metrics allow for a better understanding of the model's performance across different classes.
- **Confusion Matrix**: A confusion matrix was generated to evaluate the true positive, true negative, false positive, and false negative rates. The matrix was visualized using a *heatmap* to provide a more intuitive understanding of the model's performance.

Cross-Validation Analysis

To assess the variance in model performance, we performed *cross-validation* on the training data using *KFold cross-validation* with *5 folds*. This technique splits the training set into five subsets and trains the model five times, each time using a different fold for validation and the remaining folds for training. The *cross-validation accuracy* scores were then computed, providing a distribution of the model's performance across different data splits. This helps in understanding the model's stability and its generalization ability to new, unseen data.

Results

Exploratory Data Analysis

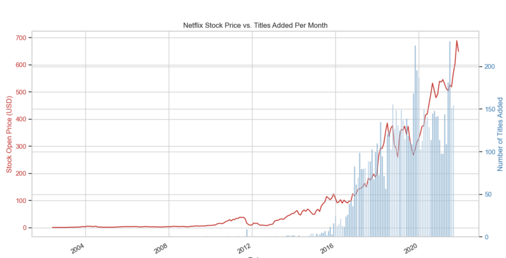

Figure 1: Netflix open stock price and number of titles added per month.

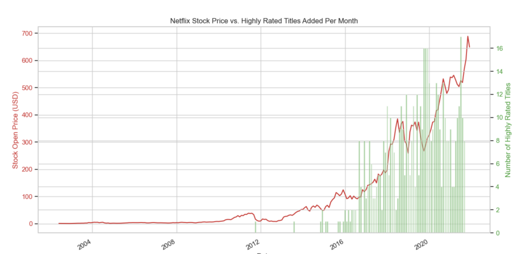

Figure 2: Netflix open stock price and number of titles with IMDB score greater than or equal to 8.0 added per month.

A qualitative data analysis shows that Netflix stock price is highly correlated with the number of titles that are added each month. In Figure 1, we see that the surge in the number of Netflix titles follows a similar pattern compared to the stock price at the start of each month. The trend is even stronger when we filter titles based on IMDB score, as seen in Figure 2.

Classification Performance

Table 1: Classification Performance of Logistic Regression Model for Netflix Stock Price Movement Prediction

Class	Precision	Recall	F1-score	Support
Down (0)	1.00	0.92	0.96	37
Up (1)	0.94	1.00	0.97	46
Accuracy		0.96		83
Macro avg	0.97	0.96	0.96	83
Weighted avg	0.97	0.96	0.96	83

The logistic regression model demonstrates exceptional performance in predicting Netflix stock price movements, achieving 96.4% accuracy on the test dataset. The model was optimized through grid search with 5-fold cross-validation across 40 parameter combinations, ultimately selecting a high regularization parameter ($C = 10000.0$) with L2 penalty and the LBFGS solver. The classification metrics reveal notable class-specific performance characteristics. For downward price movements (class 0), the model achieved perfect precision (1.00) but slightly lower recall (0.92), indicating that while all predicted downward movements were correct, approximately 8% of actual downward movements were misclassified. Conversely, for upward movements (class 1), the model attained perfect recall (1.00) with a precision of 0.94, suggesting that the model successfully identified all actual upward movements but had a small false positive rate. The balanced F1-scores for both classes (0.96 and 0.97) demonstrate that the model is equally effective at predicting both upward and downward price movements, with no significant bias toward either class despite the slightly imbalanced dataset (37 downward vs. 46 upward movements).

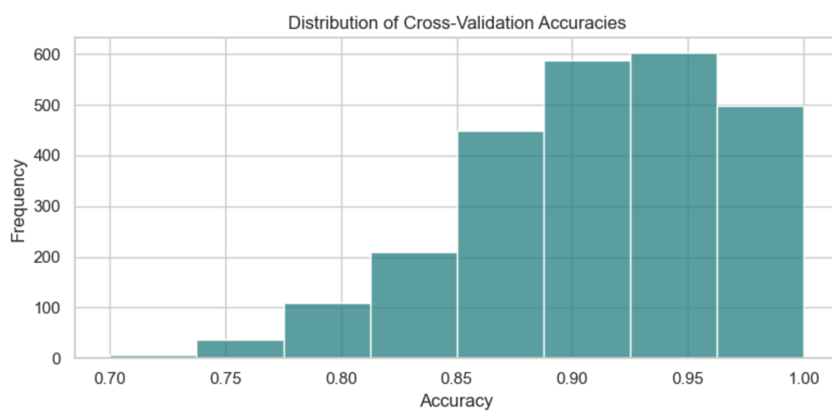

Figure 3: Distribution of cross validation accuracies for different random splits 70-30 splits of training and test data.

Table 2: Cross-Validation Accuracy Statistics (500 iterations × 5 folds)

Metric	Value
Mean Accuracy	0.9055
Standard Deviation	0.0518
Minimum Accuracy	0.7000
Maximum Accuracy	1.0000

There is some variability in the accuracy of the model depending on the choice of training data. To better understand this variance in the model, we took multiple splits of training/test data and accumulated the cross-validation accuracies of the logistic regression model. The resulting histogram in Figure 3 shows the distribution of the cross validation scores across all the training sets. The average accuracy is about 90.55%, indicating that the logistic regression model trained with monthly movie data is a very good predictor on average.

Discussion

Our findings demonstrate the remarkable effectiveness of logistic regression models in predicting Netflix stock price movements when incorporating both traditional financial indicators and content-related features. The exceptionally high test accuracy of 96.4% surpasses many previously reported models in stock market prediction literature, where accuracies typically range between 60% and 80% Sezer et al. [2020]. This performance is particularly notable given the inherent volatility and unpredictability of stock markets in general.

The integration of content quality metrics, specifically the number of high-rated titles added to Netflix's library, represents a key innovation in our approach. As shown in Figure 2, there appears to be a stronger relationship between stock prices and high-rated content additions compared to the overall volume of content (Figure 1). This aligns with consumer behavior research suggesting that content quality significantly influences subscriber retention and acquisition rates Hassouna et al. [2020], which in turn affect investor sentiment and stock performance.

The model's balanced performance across both upward and downward price movements is particularly valuable for practical applications. The perfect precision (1.00) for downward movements means that when our model predicts a decline, investors can be highly confident in this prediction. Similarly, the perfect recall (1.00) for upward movements indicates that the model captures all actual price increases, minimizing the opportunity cost of missed investment opportunities.

The cross-validation analysis provides crucial insights into the model's robustness. Although the mean accuracy of 90.55% across 500 iterations with different data splits is lower than our test set accuracy, it remains impressively high. The standard deviation of 5.18% indicates reasonable stability, though the range from 70% to 100% suggests that performance can vary depending on specific data splits. This variability underscores the importance of robust validation techniques when developing stock prediction models for real-world applications.

It is worth noting that our logistic regression model achieves this performance with relatively minimal computational complexity compared to deep learning approaches that have gained popularity in financial prediction tasks Li et al. [2016]. This demonstrates that simpler models, when properly tuned and fed with relevant features, can compete with more complex algorithms in certain prediction scenarios.

Conclusion

This study demonstrates that logistic regression models incorporating both financial indicators and content-related features can effectively predict Netflix stock price movements with high accuracy. Our findings highlight the importance of domain-specific feature engineering in financial prediction tasks, particularly for companies where product quality metrics can be quantitatively assessed.

The strong relationship between high-rated content additions and subsequent stock performance suggests that investors are responsive to indicators of Netflix's content quality, not just the volume of new releases. This insight could extend beyond Netflix to other content-driven platforms and subscription services where quality metrics are available.

Future work could explore several promising directions: (1) incorporating additional content-related features such as genre diversity or original versus licensed content ratios; (2) extending the prediction window beyond monthly movements to both shorter and longer timeframes; (3) employing more sophisticated machine learning models while maintaining interpretability; and (4) applying similar approaches to other streaming services or content platforms to test the generalizability of our findings.

In conclusion, our research demonstrates that predictive modeling of stock prices can benefit significantly from the integration of domain-specific features that capture fundamental aspects of a company's business model and value proposition. For content-driven companies like Netflix, the quality and volume of new content offerings provide valuable signals for predicting future market performance.

Acknowledgements

Thank you to Professor Jorge Silva for his support on our project. Thank you to the University of North Carolina at Chapel Hill for allowing use of their computational resources through the Longleaf cluster.

References

George S. Atsalakis and Kimon P. Valavanis. Surveying stock market forecasting techniques–Part II: Soft computing methods. *Expert Systems with Applications*, 36(3):5932–5941, 2009. doi: 10.1016/j.eswa.2008.07.006.

Michel Ballings, Dirk Van den Poel, Nathalie Hespeels, and Ruben Gryp. Evaluating multiple classifiers for stock price direction prediction. *Expert Systems with Applications*, 42(20):7046–7056, 2015. doi: 10.1016/j.eswa.2015.05.013.

Carlos A. Gomez-Uribe and Neil Hunt. The Netflix recommender system: Algorithms, business value, and innovation. *ACM Transactions on Management Information Systems*, 6(4):1–19, 2016. doi: 10.1145/2843948.

Benjamin Graham and David L. Dodd. *Security analysis: Principles and technique*. McGraw-Hill, New York, 6 edition, 2006. ISBN 9780071592536.

Mohamad Hassouna, Ali Tarhini, Tariq Elyas, and Mohammad S. AbouTrab. Customer retention in the era of digital streaming: Evidence from Netflix. *Technology in Society*, 63:101520, 2020. doi: 10.1016/j.techsoc.2020.101520.

Bruno M. Henrique, Vinicius A. Sobreiro, and Herbert Kimura. Literature review: Machine learning techniques applied to financial market prediction. *Expert Systems with Applications*, 124:226–251, 2019. doi: 10.1016/j.eswa.2019.01.012.

Gareth James, Daniela Witten, Trevor Hastie, and Robert Tibshirani. *An introduction to statistical learning*. Springer, New York, 2013. ISBN 9781461471370. doi: 10.1007/978-1-4614-7138-7.

Xiao Li, Haoran Xie, Ran Wang, Yi Cai, Jingjing Cao, Feng Wang, Huaqing Min, and Xiaojun Deng. Empirical analysis: Stock market prediction via extreme learning machine. *Neural Computing and Applications*, 27(1):67–78, 2016. doi: 10.1007/s00521-014-1550-z.

John J. Murphy. *Technical analysis of the financial markets: A comprehensive guide to trading methods and applications*. New York Institute of Finance, New York, 1999. ISBN 9780735200661.

Netflix, Inc. Annual report 2022. Annual report, Netflix Investor Relations, 2023. URL `https://ir.netflix.net/financials/annual-reports-and-proxies/`.

Jigar Patel, Sahil Shah, Priyank Thakkar, and Ketan Kotecha. Predicting stock and stock price index movement using trend deterministic data preparation and machine learning techniques. *Expert Systems with Applications*, 42(1):259–268, 2015. doi: 10.1016/j.eswa.2014.07.040.

Omer Berat Sezer, Mehmet Ugur Gudelek, and Ahmet Murat Ozbayoglu. Financial time series forecasting with deep learning: A systematic literature review: 2005–2019. *Applied Soft Computing*, 90:106181, 2020. doi: 10.1016/j.asoc.2020.106181.

David West. Neural network credit scoring models. *Computers & Operations Research*, 27(11-12): 1131–1152, 2000. doi: 10.1016/S0305-0548(99)00149-5.