Head-to-Head: Machine Learning for Influence Comparison in Social Networks

Parth Garg
2022351

Varun Kumar
2022563

Abstract

This project explores the use of machine learning to predict influential individuals within social networks. By comparing multiple models and leveraging key network features alongside human judgment data, the project aims to identify the more influential individual between two given entities. Through robust exploratory data analysis, feature engineering, and evaluation, the project provides insights into influence dynamics within social networks. The dataset, methodology, and models are available on the GitHub repository: GitHub Repository Link.

1. Motivation

Social networks play a critical role in shaping opinions, driving trends, and influencing behavior. Identifying influential individuals within these networks is essential for targeted marketing, enhancing user engagement, and understanding network dynamics. This understanding helps businesses tailor their marketing strategies, optimize content delivery, and make informed decisions. Additionally, identifying influential nodes helps in understanding network structure and dynamics, aiding in the dissemination of information and interventions. This project aims to build a robust model for predicting influential individuals in social networks, leveraging machine learning to improve strategic decision-making.

2. Related Work

1. **Kempe, D., Kleinberg, J., & Tardos, É. (2003).** Maximizing the spread of influence through a social network. *Proceedings of the Ninth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 137–146.*

   This paper presents methods for identifying key influencers in a network, providing guarantees for selecting individuals who maximize influence spread, which is crucial for improving machine learning models.

2. **Li, J., Peng, W., Li, T., Sun, T., Li, Q., & Xu, J. (2014).** Social network user influence sense-making and dynamics prediction. *Expert Systems with Applications, 41(11), 5115–5124.*

   This study introduces a framework to measure influence characteristics and proposes a dynamic model to predict influence in social networks, showing improved prediction performance using a large Twitter dataset.

3. **Utz, S. (2010).** Show me your friends and I will tell you what type of person you are: How one's profile, number of friends, and type of friends influence impression formation on social network sites. *Journal of Computer-Mediated Communication, 15(2), 314–335.*

   This research examines how friends' behaviors impact each other's actions in a microblogging network, introducing a new concept of social influence locality.

3. Dataset

3.1. Description

The dataset, which is inspired from Kaggle Link, contains Twitter activity for two individuals in each datapoint. The discrete label represents human judgment about which of the two individuals is more influential. The goal is to train a machine learning model to predict, for a pair of individuals, who is considered more influential based on the provided features.

3.2. Features

The dataset includes several features for both individuals, referred to as A and B. These features are follower count, following count, listed count, mentions received, retweets received, mentions sent, retweets sent, and the total number of posts. Additionally, the dataset includes three network-related features, labeled as Network feature 1, Network feature 2, and Network feature 3. Each of these features is provided separately for both individuals, with labels 'A' and 'B' distinguishing the first and second individuals in the pair.

4. Exploratory Data Analysis

The dataset has been processed to ensure that there are no null values. This preprocessing step is crucial for ensuring the integrity of the data. Different visualizations of the exploratory data analysis (EDA) can be viewed below:

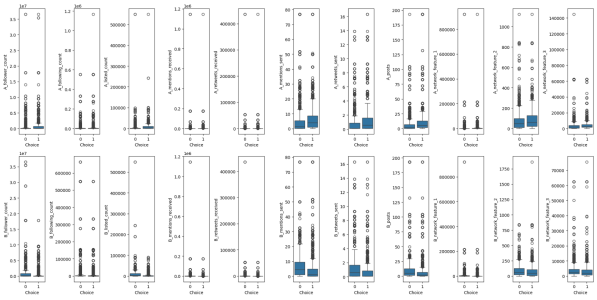

Figure 1. Box Plots of Features by Choice

Figure 1 above shows the box plots representing the distribution of each feature in relation to the 'Choice' variable, which reflects human judgment regarding the influence of individuals. Each box plot illustrates the variation of features across the two choices, indicating how specific attributes impact perceived influence.

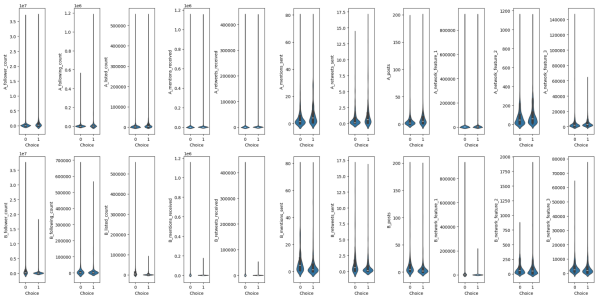

Figure 2. Violin Plots of Features by Choice

Figure 2 above displays the violin plots for each feature in relation to the 'Choice' variable, which represents human judgment regarding individual influence. The width of each violin in the plots indicates the density of the data at different values, allowing for a nuanced understanding of how the feature distributions vary between the two choices influencing human judgment on individual influence.

Figure 3 below presents the heatmap of correlation coefficients among the features in our dataset related to human influence judgments. Consequently, the following features were dropped: `A_mentions_received`, `B_mentions_received`, `A_network_feature_3`, and `B_network_feature_3`.

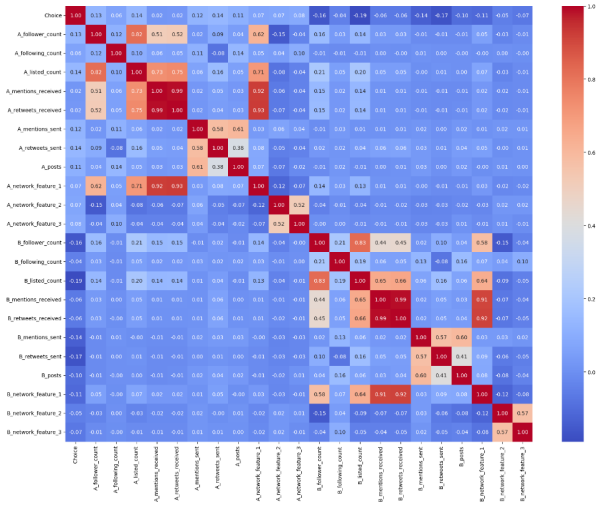

Figure 3. Correlation Heatmap of Features

These visualizations guide further analysis and feature selection in our models.

5. Methodology

This section outlines the methodology employed in our machine learning project, detailing data preparation, model selection, and evaluation processes. The goal of this study is to predict human judgment regarding influence based on various features extracted from the dataset.

5.1. Data Loading and Splitting

We begin by loading the dataset using Pandas. The dataset is divided into training and testing sets to evaluate model performance. The data is split as follows:

- **Training Data:** 80% of the dataset
- **Testing Data:** 20% of the dataset

The 'Choice' column, which serves as our target variable indicating human judgment about influence, is separated from the feature set in both training and testing datasets.

5.2. Pre-processing

5.2.1 Missing Values

The exploratory data analysis (EDA) conducted prior to modeling indicated that there were no missing values in the dataset. Thus, no additional preprocessing steps were necessary to handle missing values.

5.2.2 Feature Scaling and Selection

Since the features are likely on different scales, we employ standardization using `StandardScaler` from

`sklearn`. This transformation centers the features around zero with a standard deviation of one, which is particularly important for algorithms sensitive to the scale of the data, such as logistic regression.

5.3. Model Selection and Training

We evaluate multiple models to determine the most effective algorithm for our classification problem. We also performed **hyperparameter tuning** wherever we deemed suitable to optimize model performance. Additionally, we applied **dimensionality reduction** techniques to enhance model efficiency and interpretability. The models evaluated have been mentioned below:

5.3.1 Logistic Regression

Logistic regression, trained on standardized data and evaluated with accuracy and classification metrics, serves as a simple, interpretable baseline for influence analysis in social networks.

5.3.2 Logistic Regression with Regularization

We extend logistic regression with L1 and L2 regularization:

- **L1 Regularization:** Encourages sparsity by penalizing the absolute size of coefficients.
- **L2 Regularization:** Penalizes the squared size of coefficients, promoting weight distribution.

Both regularized models are compared to the baseline to enhance generalization.

5.3.3 ElasticNet Regularization

ElasticNet combines L1 and L2 penalties, balancing feature selection and overfitting reduction. This approach is valuable for identifying influential features.

5.3.4 Naive Bayes Classifier

The Gaussian Naive Bayes classifier is efficient and suitable for features with Gaussian distribution, making it a good choice for preliminary influence analysis.

5.3.5 Decision Tree Classifier

Decision Trees build a tree-like structure based on feature splits and can capture non-linear relationships. They are interpretable and help identify key features for influence comparison.

5.3.6 Random Forest Classifier

Random Forests build multiple decision trees and combine their results for better accuracy and reduced overfitting. This ensemble method is ideal for complex influence modeling.

5.3.7 Support Vector Machine

SVM identifies the optimal hyperplane for separating classes, excelling in high-dimensional, non-linear spaces and handling complex classification tasks.

5.3.8 Multilayer Perceptron

MLP, a type of neural network, uses multiple layers with activation functions to capture complex, non-linear relationships, making it suitable for analyzing intricate social network patterns.

5.4. Model Evaluation

For each model, we assess performance through the following metrics:

- **Accuracy:** The proportion of true results (both true positives and true negatives) among the total number of cases examined.
- **Confusion Matrix:** A summary of prediction results that shows the counts of true positive, true negative, false positive, and false negative predictions.
- **Classification Report:** Provides a detailed breakdown of precision, recall, and F1-score for each class in the dataset.

This methodology outlines the systematic approach taken to preprocess the dataset and implement various classification models to predict human judgment on influence. By employing different algorithms and evaluation techniques, we aim to identify the most effective model for our specific dataset and classification task.

6. Results

In this section, we present the results obtained from various machine learning models applied to predict human judgment about influence. The performance of each model is evaluated using accuracy, classification reports, and confusion matrices. The best accuracies for every model evaluated can be seen in the table below.

Model	Best Accuracy (%)
Logistic Regression	73.95
Logistic Regression with L1 Regularization	73.27
Logistic Regression with L2 Regularization	73.55
Logistic Regression with ElasticNet Regularization	73.95
Gaussian Naive Bayes	62.45
Decision Tree	76.41
Random Forest	77.93
Multilayer Perceptron	75.84
Support Vector Machine	74.18

Table 1. Summary of Models and Their Best Accuracy

6.1. Comparison of Models

The comparison of the various models indicates that the Random Forest classifier outperforms other models, achieving an accuracy of 77.93%. The Decision Tree model also demonstrated strong performance, with an accuracy of 76.41% after hyperparameter tuning, showcasing its ability to capture complex patterns. On the other hand, the Gaussian Naive Bayes model underperformed significantly, with an accuracy of 62.45% even after dimensionality reduction, suggesting it is not well-suited for this dataset due to its simplistic assumptions about feature independence.

These results highlight the importance of selecting appropriate algorithms based on data characteristics. Random Forest's ensemble learning mitigated overfitting effectively, while Decision Tree's interpretability and flexibility made it a strong contender. However, Gaussian Naive Bayes struggled with this dataset's complexity, emphasizing that algorithm choice must align with the underlying data structure. The findings also underscore the role of techniques like hyperparameter tuning and dimensionality reduction in enhancing model performance.

6.2. Discussion

In conclusion, the comprehensive analysis of various machine learning models on the social network influence prediction task revealed that the Random Forest Classifier performed the best in terms of accuracy, precision, and overall robustness. This model's ensemble approach handles complex, non-linear relationships and interactions between features, making it suitable for understanding the nuanced nature of influence in social networks. Its ability to aggregate decision trees ensures stability and generalizability, making it a powerful tool for capturing influence dynamics. Understanding influence in social networks is crucial as it empowers targeted marketing and informs strategic decision-making, ultimately driving user engagement.

7. Timeline

We were able to follow our proposed timeline:

- **Week 0:** Proposal
- **Week 1:** Problem Framing and Data Gathering
- **Weeks 2-3:** Data Preprocessing and Exploratory Data Analysis (EDA)
- **Weeks 4-5:** Feature Engineering and Selection
- **Weeks 6-8:** Model Training and Evaluation
- **Weeks 9-10:** Model Deployment, Testing, and Optimization

8. Team Contributions

- **Parth Garg:** Documentation and Report Writing, Decision Tree, Random Forest, Multilayer Perceptron, Support Vector Machine, Feature Selection, Results and Analysis
- **Varun Kumar:** Documentation and Report Writing, Exploratory Data Analysis, Logistic Regression and Regularizations, Naive Bayes, Dimensionality Reduction, Hyperparameter Tuning, Results and Analysis