**Project 2 - Literature Study Using Machine Learning**

Jessica Taylor

Department of Computer Science, Charleston Southern University

CSCI 409: Fundamentals of Artificial Intelligence

Dr. Yue

November 25, 2024

**1. Introduction**

The goal of this project was to explore the relationship between research paper abstracts and their citation counts. Using machine learning techniques, I sought to predict citation counts based on abstract text features. Two regression models, Random Forests and Linear Regression, were implemented for this task.

**2. Hypothesis Development and Feature Selection**

**Hypothesis**

The citation count of a research paper is shaped by the content and linguistic features of its abstract, as the abstract serves as a concise summary highlighting the paper's contributions, originality, and importance.

**Feature Selection/Generation**

To test this hypothesis, I focused on the textual data from the Abstract column. Text was transformed into numerical features using TF-IDF Vectorization, a popular method for text analysis:

* **Why TF-IDF?** It emphasizes important terms while reducing the influence of commonly used words (e.g., "the," "and") that do not contribute much meaning.
* **Parameters**:
  + Limited to the top 5000 features to reduce computational complexity.
  + Removed common English stop words to focus on domain-relevant terms.

In addition, citation counts were log-transformed to handle their skewed distribution. This ensured the target variable was more evenly distributed, making it easier for models to learn patterns.

**3. Models**

I chose Random Forest Regressor and Linear Regression for their contrasting characteristics and potential to provide meaningful insights and perform well.

**Why Random Forest?**

Random Forest is a robust ensemble learning method that aggregates the results of multiple decision trees:

1. **Non-linear Relationships**: It can capture non-linear patterns between text features and citation counts.
2. **Feature Importance**: Random Forest provides insights into the importance of features, which can help interpret results.
3. **Robustness**: It handles high-dimensional datasets well and is resilient to overfitting due to its ensemble nature.

**Why Linear Regression?**

Linear Regression serves as a baseline model to compare against more complex approaches:

1. **Simplicity**: It assumes a linear relationship between features and the target variable, making it easy to interpret.
2. **Benchmarking**: Linear regression provides a straightforward comparison to gauge the performance of more advanced models.
3. **Efficiency**: It is computationally efficient for initial experimentation.

**4. Experimental Setup**

* **Dataset**: I use the dataset provided on Blackboard, which contains paper abstracts and citation counts.
* **Data Cleaning**:
  + Removed missing values in the Abstract and Article Citation Count columns. There are no missing values in the Blackboard dataset; however, most datasets do contain missing values and these need to be handled.
  + Applied logarithmic transformation to citation counts for stability.
* **Train-Test Split**: Data was split into 80% for training and 20% for testing to evaluate model performance.
* **Evaluation Metrics**:
  + **RMSE (Root Mean Square Error)**: Measures the average error magnitude.
  + **R2R^2R2 Score**: Indicates how well the model explains variance in citation counts.

**5. Results**

The performance of the models is summarized below:

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| |  |  |  | | --- | --- | --- | | **Model** | **RMSE** | **R2R^2R2 Score** | | Random Forest | 1.19 | -0.07 | | Linear Regression | 1.21 | -0.09 | |  |
|  |  |  |

**6. Observations**

* **Performance**:
  + Both models exhibited poor predictive performance, as indicated by negative R2R^2R2 scores.
  + Random Forest outperformed Linear Regression slightly, achieving a lower RMSE.
* **Visualization**:
  + The scatter plots (Figure 1) of true vs. predicted values revealed significant deviations from the ideal diagonal line, confirming that the predictions were not well-aligned with actual values.

**Challenges Identified**

* The textual features from abstracts alone might not capture the full complexity of factors influencing citation counts.
* Citation counts are affected by external factors (e.g., journal prestige, author network) that were not included in the dataset.

**7. Recommendations**

To improve performance:

1. **Expand Feature Set**:
   * Include metadata like year of publication, journal impact factor, and author details.
   * Analyze linguistic complexity (e.g., readability scores) and sentiment analysis of abstracts.
2. **Dataset Expansion**:
   * Use a larger and more diverse dataset to improve model generalization.

**8. Conclusion**

This project explored the use of machine learning models to predict citation counts from research paper abstracts. While both Random Forest and Linear Regression models struggled to accurately predict citation counts, the results highlighted the need for additional features and dataset expansion. Random Forests slight edge demonstrates its ability to capture more complex patterns compared to Linear Regression.

This study forms a foundation for future work, with improvements likely achievable through better feature engineering and use of larger datasets.

**9. Figures**

A comparison of a graph

Description automatically generated

**Figure 1: Performance Comparison – True v. Predicted Citation**