A DUAL-TASK STUDY: LIFE EXPECTANCY PREDICTION AND SENTIMENT ANALYSIS OF FILM REVIEWS

Zili Gong, Jihan Li, Chunlin Wang, Qijun Han

School of Automation and Intelligent Manufacturing, Southern University of Science and Technology
Shenzhen, Guangdong, China
{gongzl2022, lijh2022, wangcl2022, hanqj2022}@mail.sustech.edu.cn

ABSTRACT

This paper presents a comprehensive study across two distinct machine learning domains. The first task focuses on predictive modeling, where we develop a model to forecast national life expectancy using a range of socio-economic and environmental indicators from 2008 to 2018. We explore various regression techniques, feature importance, and model improvement strategies. The second task delves into natural language processing, conducting sentiment analysis on Douban movie reviews. We implement and compare traditional machine learning classifiers with modern Large Language Model (LLM) approaches, evaluating their effectiveness in discerning positive from negative sentiment in textual data. This work highlights the application of diverse statistical methods to solve real-world prediction and classification problems.

***Index Terms*—** Life Expectancy, Predictive Modeling, Sentiment Analysis, Machine Learning, Natural Language Processing, LLM

1. INTRODUCTION

This report details our work on two data science projects. The first project, "Life Expectancy," involves predicting life expectancy at birth based on 12 features for 211 countries. The primary objective is to train a model on data from 2008-2017 to predict life expectancy for the year 2018, using the *life_indicator_2008-2018* dataset.

The second project, "Douban Movie Comment Analysis," aims to classify the sentiment of movie reviews from Douban as either positive or negative. This task utilizes the *douban_movie* dataset. We explore both traditional machine learning techniques and the capabilities of Large Language Models (LLMs) for this text classification problem.

2. TASK 1: LIFE EXPECTANCY PREDICTION

The goal of this task is to build a regression model to predict 'Life expectancy at birth' using various national indicators.

2.1. Data Understanding

The dataset contains 12 features, including 'Agriculture, forestry, and fishing, value added (

A correlation heatmap was generated to visualize the relationships between features. Missing data was a significant issue, and we compared several imputation methods, including mean/median filling, interpolation, and K-Nearest Neighbors (KNN) imputation.

TODO: Insert the correlation heatmap figure. Discuss which imputation method was chosen and why, based on performance comparisons.

2.2. Modeling

We trained and evaluated several regression models to identify the best predictor for life expectancy. The models included Linear Regression, Lasso, Ridge, Random Forest, XGBoost, and Support Vector Regression (SVR). The data from 2008 to 2017 served as the training set, and the 2018 data was used for testing.

Model performance was evaluated using Mean Squared Error (MSE) and the coefficient of determination (R^2).

TODO: Present a table comparing the MSE and R^2 scores for each model on the 2018 test set. Analyze the results and select the best-performing model.

Feature importance was extracted from the best models (e.g., coefficients from linear models, feature importance scores from tree-based models) to identify the key drivers of life expectancy.

TODO: List the top 5 most important features and discuss whether they align with the initial hypotheses.

2.3. Analysis of Predictions

We visualized the residuals (the difference between predicted and actual values) for the 2018 data to assess the model's accuracy. Outliers, i.e., countries where the prediction error was particularly large, were identified.

TODO: Include a plot of predicted vs. actual values for 2018. Identify any major outliers and provide potential ex-

planations for the large prediction errors (e.g., unique socio-economic events in those countries in 2018). Analyze the distribution of prediction errors.

2.4. Model Improvement

To enhance model performance, we employed stepwise forward selection to find an optimal subset of features. Additionally, we engineered new features, such as 'GDP per capita' (GDP / Population), to better capture the economic status of a country.

TODO: Describe the results of the model improvement techniques. Did stepwise selection or feature engineering lead to a significant improvement in MSE or R^2?

3. TASK 2: DOUBAN MOVIE COMMENT ANALYSIS

This task focuses on binary sentiment classification of movie reviews. Reviews with star ratings of 1 or 2 were labeled as negative, while those with ratings of 3, 4, or 5 were labeled as positive.

3.1. Part 1: Machine Learning Approach

3.1.1. Text Preprocessing

The raw text comments were preprocessed to prepare them for vectorization. This involved tokenization (using a Chinese tokenizer like Jieba), removal of stopwords, special symbols, and low-frequency words.

3.1.2. Text Vectorization

We converted the cleaned text into numerical vectors using TF-IDF.

TODO (Bonus): If Word2Vec or BERT embeddings were used, describe the process and compare the results with TF-IDF.

3.1.3. Model Training & Evaluation

We trained and cross-validated Logistic Regression and Naive Bayes classifiers on an 80/20 train/test split of the data. Performance was measured using accuracy, precision, recall, and F1-score.

TODO: Present a table with the evaluation metrics for both models. Analyze their performance and discuss their respective strengths and weaknesses for this task.

3.2. Part 2: Large Language Model (LLM) Approach

3.2.1. Prompt Design & In-Context Learning

To leverage LLMs for this task, we designed effective prompts. We experimented with zero-shot and few-shot learning. For few-shot learning, the prompt included examples of positive and negative reviews to guide the model. An example of a few-shot prompt structure is:

```
Based on the examples, determine the sentiment of

1. Comment: The acting was superb, and the plot wa
   Sentiment: Positive

2. Comment: The special effects were terrible and
   Sentiment: Negative

Now analyze:
Comment: {{review_text}}
Sentiment:
```

3.2.2. LLM API Testing

We used the APIs for two LLMs (e.g., ChatGPT-3.5, DeepSeek) to predict the sentiment of a sample of reviews from our test set.

TODO: Name the specific LLMs used.

3.2.3. Discussion

We compared the performance of the traditional machine learning models with the LLM-based predictions.

TODO: Compare the accuracy of the LLMs against the Logistic Regression and Naive Bayes models. Discuss the strengths and limitations of each approach. Analyze specific cases where the ML and LLM predictions differed and provide insights into why.

4. BONUS TASKS

4.1. Task 1 Bonus: Forecasting to 2025

A key challenge explored was the feasibility of forecasting life expectancy for 2025. This requires extrapolating the feature trends from 2008-2018 and feeding them into the trained regression model.

TODO: Discuss the methodology used for feature extrapolation (e.g., time series forecasting on each feature) and present the 2025 life expectancy predictions. Analyze the confidence and potential error sources of this long-range forecast.

4.2. Task 2 Bonus: Advanced NLP Exploration

4.2.1. Advanced Data Analysis

We conducted further analysis on the review data. This included generating word clouds to visualize the most frequent terms in positive and negative reviews and exploring the correlation between review length and the assigned star rating.

TODO: Insert word cloud visualizations and a plot showing the relationship between review length and rating. Discuss any insights gained.

4.2.2. LLM Prompt Optimization

To improve LLM performance, we tested more sophisticated prompting strategies, such as chain-of-thought and role-playing prompts, analyzing how changes in prompt phrasing affected the quality of sentiment prediction.

TODO: Provide examples of the advanced prompts used and compare their performance to the initial few-shot prompts.

4.2.3. Fine-Tuning LLMs

We fine-tuned an open-source LLM (e.g., LLaMA-2) on a small subset of the Douban review data. The performance of the fine-tuned model was then compared against its pre-trained counterpart to evaluate the effectiveness of domain-specific adaptation.

TODO: Describe the fine-tuning process and present a comparison of the model's performance before and after fine-tuning.

5. COMPLIANCE WITH ETHICAL STANDARDS

This research study [1] was conducted using publicly available data. The life expectancy data is aggregated at a country level, and the movie review data is anonymized. Therefore, no formal ethics approval was required for this study.

6. ACKNOWLEDGMENTS

No external funding was received for conducting this study. The authors have no relevant financial or non-financial interests to disclose.

7. REFERENCES

[1] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner, "Gradient-based learning applied to document recognition," *Proceedings of the IEEE*, vol. 86, no. 11, pp. 2278–2324, 1998.

[2] L. van der Maaten and G. Hinton, "Visualizing data using t-SNE," *Journal of Machine Learning Research*, vol. 9, pp. 2579–2605, 2008.