Internship Final Report

Student Name: Ashish Mishra
University: University of Mumbai
Major: Electronics Computer Science
Internship Duration: April 1th, 2025 - May 5th, 2025
Company: ShadowFox
Domain: AI/ML
Coordinator: Mr. Aakash

Objectives

1. Master data preprocessing, exploratory data analysis (EDA), and machine learning techniques for regression and natural language processing (NLP) tasks.

2. Gain proficiency in Python libraries (Pandas, Scikit-learn, Matplotlib, Seaborn, Hugging Face) and tools like Git for professional workflows.

3. Build, evaluate, and analyze predictive models and language models to address real-world problems.

Tasks and Responsibilities

1. **Beginner Task: Boston House Price Prediction**
    - Description: Developed a regression model to predict Boston house prices using a dataset with features like number of rooms (RM) and crime rate (CRIM).
    - Tasks:
        - Preprocessed the dataset (506 entries, 14 features) by handling missing values in columns like CRIM and ZN using Simple Imputer (median strategy) and scaling features with Standard Scaler.
        - Conducted EDA using Seaborn and Matplotlib, creating histograms and correlation heatmaps to identify key predictors (e.g., RM, LSTAT).
        - Trained a Linear Regression model, splitting data into 80% training and 20% testing sets.
        - Evaluated performance with Mean Squared Error (MSE) of 24.983, Root Mean Squared Error (RMSE) of 4.998, and R-squared of 0.659.
        - Visualized actual vs. predicted prices using scatter plots.

2. **Intermediate Task: Car Selling Price Prediction and Analysis**
    - Description: Built an ML model to predict car selling prices based on features like fuel type, years of service, and kilo meters driven.
    - Tasks:
        - Preprocessed the dataset (301 entries) by deriving Years_Used and encoding categorical variables (Fuel_Type, Seller_Type, Transmission) with one-hot encoding.
        - Performed EDA with histograms and correlation heatmaps, noting strong correlations between Selling_Price and Present_Price.

        - Trained a Random Forest Regressor, optimizing hyperparameters via Randomized Search CV.
        - Evaluated performance using MSE and RMSE (specific values documented in the Notebook).
        - Visualized feature distributions and correlations for insights.

3. **Advanced Task: AI-Driven Natural Language Processing Project**
    - **Description:** Implemented and analyzed a language model (e.g., BERT or a Hugging Face transformer) to explore NLP capabilities, as outlined in the advanced task requirements.
    - **Tasks:**
        - Selected a pre-trained language model for tasks like text classification, sentiment analysis, or text generation.
        - Implemented the model in a Jupyter Notebook, setting up the pipeline using Hugging Face's Transformers library.
        - Conducted exploratory analysis on sample text inputs to evaluate contextual understanding, response quality, and limitations.
        - Visualized results, such as attention mechanisms or performance metrics (e.g., accuracy, F1-score), using Matplotlib or Plotly.
        - Defined research questions to investigate strengths (e.g., contextual accuracy) and weaknesses (e.g., handling ambiguous inputs).

Learning Outcomes

1. **Technical Proficiency:** Mastered Python libraries (Pandas, NumPy, Scikit-learn, Matplotlib, Seaborn) for regression tasks and Hugging Face for NLP, applying them to diverse problems.

2. **Data Analysis and Modeling:** Gained expertise in EDA, feature engineering, and model evaluation, using metrics like R-squared, RMSE, and NLP-specific metrics (e.g., F1-score).

3. **NLP Expertise:** Learned to implement and analyze language models, understanding their capabilities in contextual processing and text generation.

4. **Professional Tools:** Improved proficiency in Git for version control and Jupyter Notebooks for clear, reproducible documentation.

5. **Time Management:** Balanced academic and internship commitments, leveraging the extended deadline to deliver high-quality work.

Challenges and Solutions

1. Challenge: Missing Values in Boston Housing Dataset
    - Issue: Missing values in columns like CRIM and ZN risked model inaccuracies.
    - Solution: Applied Simple Imputer with a median strategy, validating post-imputation distributions to ensure data integrity.
2. Challenge: Overfitting in Car Price Prediction
    - Issue: Initial Random Forest models showed overfitting, with high training accuracy but lower test performance.
    - Solution: Used RandomizedSearchCV to optimize hyperparameters (e.g., number of trees, max depth) and cross-validation to improve generalization.
3. Challenge: Complexity of Language Model Implementation
    - Issue: Setting up and analyzing a language model was complex due to limited prior NLP experience.
    - Solution: Studied Hugging Face tutorials and documentation to configure the model pipeline. Experimented with pre-trained models and sought mentor feedback to refine the approach.
4. Challenge: Balancing Academic and Internship Deadlines
    - Issue: Overlapping exams created time constraints for completing three tasks.
    - Solution: Created a prioritized schedule and utilized the extended deadline (May 5th, 2025) to submit all deliverables, including Notebooks and POW videos, on time.

Conclusion

My internship at ShadowFox was a transformative journey that bridged theoretical AI/ML knowledge with practical expertise across regression and NLP. The Boston House Price Prediction task built my foundation in data preprocessing and Linear Regression, while the Car Selling Price Prediction task deepened my skills in feature engineering and Random Forest modeling. The advanced NLP project expanded my understanding of language models and their real-world applications. Overcoming challenges like missing data, overfitting, and NLP complexity strengthened my problem-solving abilities and technical confidence. The guidance from Mr. Hariharan and Mr. Aakash was pivotal in navigating project requirements and ensuring timely submissions. This experience has equipped me with industry-relevant skills and a professional mindset, preparing me for future roles in AI and data science.

Acknowledgments

I extend my heartfelt gratitude to ShadowFox for providing this opportunity to grow as an AI/ML professional. My mentor, Mr. Hariharan, offered invaluable guidance and feedback that shaped my project outcomes. I thank Mr. Aakash, the internship coordinator, for his support in clarifying submission processes and extending deadlines to accommodate academic commitments.