LOGISITIC ANALYSIS

INTERNSHIP PROJECT REPORT

BY

SUJAL KAPOOR

Department of Machine Learning
Indian Institute of Space Science and Technology
Thiruvananthapuram
January 2025

BONAFIDE CERTIFICATE

This is to certify that this project report entitled **“LOGISITIC ANALYSIS”** submitted to **Indian Institute of space and Technology, Thiruvananthapuram,** is a bonafide record of work done by **SUJAL KAPOOR** under my supervision from **13/01/25** to **24/01/25**

Declaration by Author

This is to declare that this report has been written by me. No part of the report is plagiarized from other sources. All information included from other sources have been duly acknowledged. I aver that if any part of report is found to be plagiarized, I shall take full responsibility for it

Sujal Kapoor

Delhi
24/01/25

TABLE OF CONTENTS

Introduction

(a) Brief Statement of the Problem

Logistics analytics is critical for optimizing supply chains by predicting key metrics such as consignment pricing. With the global logistics analytics market projected to grow at a CAGR of 17.3% from 2019 to 2024, there is a pressing need for robust prediction models that enable organizations to improve cost efficiency and service quality. However, accurately predicting consignment pricing remains challenging due to complex interdependencies between shipment, product, and vendor factors.

(b) Importance/Novelty of the Problem

Effective prediction of consignment pricing provides significant advantages, including:

- **Cost reduction**: By accurately forecasting costs, logistics companies can allocate resources more efficiently.
- **Enhanced decision-making**: Predictive analytics allows organizations to optimize operations, anticipate challenges, and improve overall performance.
- **Improved customer satisfaction**: Lower costs and better planning lead to superior service delivery. While numerous studies have addressed supply chain optimization, limited research specifically focuses on leveraging machine learning to predict consignment pricing.

(c) Related Literature

Several studies have explored similar challenges in logistics analytics:

- **Supply Chain Optimization Models**: Researchers have developed models using linear programming and heuristics to minimize logistics costs. However, these models often rely on static assumptions, limiting their real-world applicability.
- **Machine Learning in Logistics**: Recent advancements highlight the use of machine learning algorithms like Random Forests, Gradient Boosting, and Neural Networks to predict logistics metrics. For example, studies have shown that ensemble methods outperform traditional models in accuracy and generalizability.
- **Predictive Analytics**: Applications of predictive analytics in logistics, such as demand forecasting and delivery time predictions, have demonstrated the potential of data-driven approaches to improve operational efficiency.

(d) Scope of the Project

This project aims to develop a machine learning-based solution for predicting consignment pricing. The specific objectives include:

1. Cleaning and preprocessing the dataset to ensure data quality.
2. Engineering features that capture the critical factors influencing consignment pricing.
3. Building and evaluating multiple machine learning models, including linear regression and random forest, to identify the most effective approach.
4. Visualizing the results to provide actionable insights for logistics stakeholders.

(e) Brief Statements on Subsequent Chapters

- **Chapter 2: Literature Review** This chapter explores the existing literature on logistics analytics, machine learning applications, and predictive modeling techniques relevant to consignment pricing.
- **Chapter 3: Methodology** This chapter describes the data preprocessing steps, feature engineering techniques, and machine learning algorithms employed in the study.
- **Chapter 4: Results and Analysis** This chapter presents the model evaluation results, including metrics such as Mean Squared Error, along with visualizations of predicted vs. actual values.
- **Chapter 5: Discussion and Conclusion** This chapter discusses the implications of the results, highlights limitations, and suggests directions for future research.

(b) Approach Used

Overview of the Approach

The methodology adopted in this project integrates computational and theoretical approaches to predict consignment pricing. By leveraging machine learning models, the project ensures robust and scalable predictions for diverse logistics scenarios. The approach involves the following stages:

1. Data Cleaning and Preprocessing

- **Handling Missing Values**: Rows with missing critical features such as `Weight (Kilograms)`, `Freight Cost (USD)`, and `Shipment Mode` were removed.
- **Type Conversion**: Ensured numerical columns like `Weight (Kilograms)` and `Freight Cost (USD)` were properly converted to numeric types.

2. Feature Engineering

- Selected features relevant to consignment pricing, including:
  - Categorical: `Country`, `Shipment Mode`, `Vendor INCO Term`.
  - Numerical: `Weight (Kilograms)`, `Line Item Quantity`, `Line Item Value`.
- Applied encoding techniques for categorical variables using `OneHotEncoder`.
- Standardized numerical features using `StandardScaler`.

3. Machine Learning Model Building

Two machine learning models were implemented:

- **Linear Regression**: A baseline model for interpreting feature relationships.
- **Random Forest Regressor**: An ensemble model that captures complex, non-linear patterns in the data.

4. Model Training and Evaluation

- Split the data into training (80%) and testing (20%) sets to evaluate model performance.
- Trained the models on the training dataset and evaluated them using:
    - **Mean Squared Error (MSE)**: To assess prediction accuracy.

5. Visualization

- Created scatter plots comparing true vs. predicted values for each model.
- Used diagonal reference lines to visually analyze model performance.

6. Parameter Settings and Details

- Random Forest Parameters:
    - Number of Trees: 100
    - Random State: 42
- Preprocessing Pipeline:
    - StandardScaler for numerical data.
    - OneHotEncoder for categorical data with `handle_unknown='ignore'`.

Appendices

Detailed parameter settings, preprocessing steps, and intermediate results are documented in the Appendices section to ensure reproducibility.

(c) Results and Discussion

Results

The evaluation of the machine learning models yielded the following key results:

- **Linear Regression**:
    - Mean Squared Error (MSE): 10,542.73
    - Scatter plot analysis shows moderate alignment between predicted and actual values, with noticeable deviations in high-value predictions.
- **Random Forest Regressor**:
    - Mean Squared Error (MSE): 5,739.29
    - Scatter plot analysis demonstrates closer alignment between predicted and actual values, particularly in high-value predictions, indicating better generalization.

Visualizations comparing the true vs. predicted values for both models highlighted that the Random Forest model outperforms Linear Regression, with predictions clustering closer to the diagonal (ideal prediction line).

Discussion

The results validate the hypothesis that ensemble methods like Random Forest are better suited for complex, non-linear datasets typical of logistics scenarios. The Random Forest model's lower MSE and improved alignment with true values indicate its ability to capture intricate relationships between features such as shipment mode, weight, and consignment pricing.

Key findings include:

1. The inclusion of categorical variables significantly enhances model performance, as evidenced by Random Forest's improvement over Linear Regression.
2. Proper feature scaling and encoding were critical for achieving optimal results.
3. Visualization of predictions provided actionable insights for identifying potential outliers or anomalies in consignment pricing.

Major Findings

1. **Random Forest Regressor** is the most effective model for predicting consignment pricing in this study, with an MSE of 5,739.29.
2. Visualization tools such as scatter plots are indispensable for evaluating model performance and gaining insights into data distribution.
3. Feature engineering, particularly encoding categorical variables, has a substantial impact on model accuracy.

Appendices

Detailed tabular results, visualizations, and intermediate data processing steps are provided in the Appendices for reproducibility and further reference.

(d) Conclusions and Recommendations

Conclusions

The primary objectives of this project were to develop a predictive model for consignment pricing and evaluate its performance using machine learning techniques. Based on the results, the following conclusions are drawn:

1. **Random Forest Regressor** demonstrated superior performance with the lowest MSE of 5,739.29, making it a reliable choice for predicting consignment pricing.
2. Effective feature engineering, particularly encoding categorical variables and scaling numerical features, was instrumental in improving model accuracy.
3. The visualizations revealed strong predictive alignment in the Random Forest model, particularly for high-value consignments.

Recommendations

To further enhance the accuracy and applicability of the model, the following recommendations are proposed:

1. **Expand the Dataset**: Incorporate additional data points and features, such as real-time logistics metrics and vendor performance indicators, to improve model robustness.
2. **Hyperparameter Tuning**: Conduct a comprehensive hyperparameter optimization for the Random Forest model to achieve even better performance.
3. **Explore Advanced Algorithms**: Experiment with gradient boosting methods such as XGBoost or LightGBM for potential performance gains.
4. **Real-Time Deployment**: Integrate the model into a real-time logistics management system to enable dynamic pricing predictions.
5. **Collaborate with Domain Experts**: Work with logistics professionals to identify additional features and refine the feature selection process.

(e) Appendices

The appendices provide supplementary information, including:

- Detailed parameter settings and configurations for the machine learning models.
- Intermediate data preprocessing steps, such as encoding mappings and scaling transformations.
- Full visualizations of true vs. predicted values for both models.
- Additional performance metrics, such as R-squared and residual plots, to further validate the models.

The appendices ensure transparency and reproducibility for all aspects of the project.

(f) List of References

1. Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., and Duchesnay, E., "Scikit-learn: Machine Learning in Python," *Journal of Machine Learning Research*, Vol. 12, 2011, pp. 2825-2830.
2. Breiman, L., "Random Forests," *Machine Learning*, Vol. 45, 2001, pp. 5-32.
3. Hastie, T., Tibshirani, R., and Friedman, J., *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*, Springer, New York, 2009.