MGT 256

Project Report

The Road to Better Deliveries: Pruning Lead Time

Group 9:

- Dev Kotian

- Gunjan Kale

- Sunny Khade

- Nidhi Verhani

- Parv Maru

Dataset: Logistics

Abstract

Minimizing delivery lead time is a significant challenge in logistics operations. This project investigates the feasibility of predicting delivery lead time using data-driven methodologies. By preprocessing and analyzing a comprehensive logistics dataset, creating relevant dummy variables, and employing advanced machine learning techniques, the team built and compared multiple predictive models. The Exploratory Data Analysis (EDA) phase provided critical insights into the dataset, guiding the team's decision-making throughout the modeling process. Variable selection, utilizing methods such as Best Subset Selection and Forward Selection, identified 13 significant factors influencing delivery lead time. The team then developed Linear Regression and K-Nearest Neighbors (KNN) models, with the Linear Regression model emerging as the most effective, achieving an impressive adjusted R-squared value of 0.9323 and a low Root Mean Squared Error (RMSE) of 0.542. This outperformance of the KNN model highlights the Linear Regression model's superior predictive accuracy and interpretability. The findings of this study demonstrate the immense potential for data analytics to optimize operational efficiency in the logistics industry. Additionally, the project underscores the crucial importance of meticulous data preprocessing and feature engineering in building robust and reliable predictive models.

Introduction

Research Problem:

The logistics industry is the backbone of global commerce, with timely and efficient deliveries playing a critical role in customer satisfaction and operational success. However, as the volume and complexity of deliveries increase, optimizing delivery lead times has become a significant challenge for logistics providers. Companies now aim not only to minimize delays but also to anticipate potential disruptions to improve service quality and maintain a competitive edge. This project focuses on leveraging machine learning and statistical modeling techniques to predict delivery lead times, with the goal of providing actionable insights to logistics managers and decision-makers.

Research Problem The primary research questions underlying this project are:

1. Can we accurately predict the lead time of deliveries in order to minimize it?
2. What are the key factors that significantly influence delivery lead time?

Project Objective:

The primary objective of this study is to identify the critical factors impacting delivery lead times and develop predictive models that can accurately forecast these lead times. By leveraging a comprehensive dataset containing various transactional and operational details, the study aims to provide actionable insights to logistics managers and decision-makers. These insights can help streamline operations, optimize resource allocation, and reduce costs, ultimately enhancing the overall efficiency of logistics operations.

Broader Implications:

This research extends beyond academic interest, as the ability to accurately predict delivery lead times can have significant real-world implications. By forecasting lead times more accurately, businesses can improve their planning and scheduling processes, mitigate the impact of supply chain disruptions, and enhance customer trust and satisfaction. Moreover, the findings of this project highlight the importance of integrating data-driven approaches into modern logistics management practices, paving the way for more data-informed decision-making in the industry.

Methodology

Data Preprocessing:

The dataset underwent comprehensive preprocessing to ensure data quality and reliability:

- **Handling Missing Values:** Missing values were identified and removed to maintain data integrity. Rows with incomplete data were omitted using the na.omit() function in R. This ensured that subsequent analyses were not biased or distorted by incomplete information.
- **Outlier Elimination:**
    - The Interquartile Range (IQR) method was applied to detect and remove outliers.
    - Negative delivery lead time values, which were logically incorrect, were excluded from the dataset. Boxplots were created to visualize distributions and identify anomalies in features like delivery_lead_time and distance_km.
- **Feature Removal:** Irrelevant columns such as id, factory_location, and destination were excluded to streamline the dataset and reduce noise in the model-building process.

Feature Engineering:

1. **Dummy Variables:** Categorical variables were converted into dummy variables using the as.factor() method. This allowed the inclusion of qualitative data such as transport_mode, shipment_urgency, and weather_conditions in the analysis.
2. **Feature Scaling:** Continuous variables were standardized to ensure uniformity across the dataset, a critical step for algorithms sensitive to variable scaling, such as KNN.

Data Partitioning:

The dataset was divided into subsets to ensure robust model evaluation:

- **60% Training Data:** Used to train the models and determine the optimal parameters.
- **40% Validation Data:** Employed to evaluate model performance and generalizability.

```
train_indices <- sample(1:nrow(logistics_cleaned),
                        size=nrow(logistics_cleaned)*0.6,
                        replace=F)

logistics_train <- logistics_cleaned[train_indices,]
logistics_validation <- logistics_cleaned[-train_indices,]
```

This partitioning strategy ensured that the models were not overfitted and could generalize to unseen data.

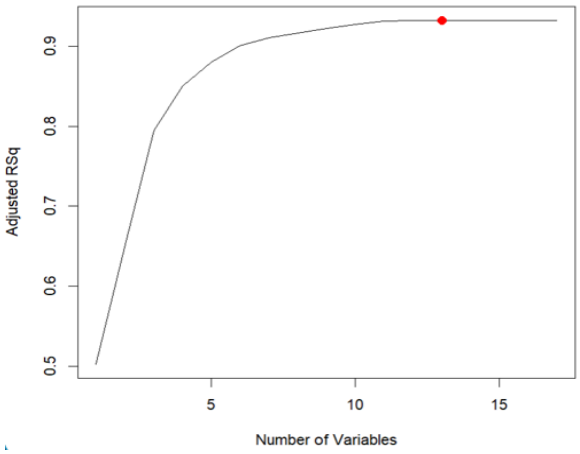
Exploratory Data Analysis (EDA)

The Exploratory Data Analysis (EDA) phase provided critical insights into the dataset, facilitating informed decision-making throughout the modeling process. The team analyzed numerical and categorical variables using a variety of visualization techniques, including histograms, box plots, and bar plots. These visualizations helped the team identify the distribution, range, and outliers within the data. Additionally, the team constructed correlation matrices to explore the relationships and interdependencies among the different variables in the dataset. Beyond the basic visualization techniques, the team also employed more advanced data exploration methods, such as clustering and time-series analysis, to uncover deeper patterns and trends within the data.

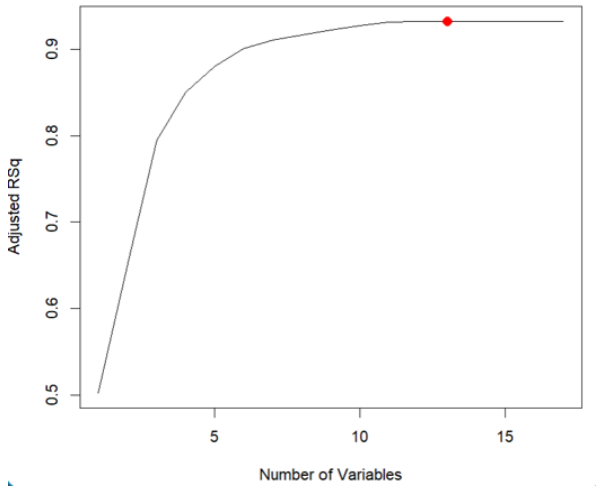
Variable Selection

Variable selection was a crucial step in the modeling process, as it involved identifying the most influential factors that impact delivery lead time. The team utilized two complementary methods for variable selection: Best Subset Selection and Forward Selection. Best Subset Selection evaluated all possible combinations of the predictor variables, while Forward Selection sequentially added the most significant variables to the model. Both of these techniques identified 13 variables that collectively maximized the adjusted R-squared value, indicating that these 13 variables were the most important in predicting delivery lead time.

Variable Selection - Best Subset Selection-



Variable Selection - Forward Selection



Modeling

The team employed two primary modeling techniques to predict delivery lead time: Linear Regression and K-Nearest Neighbors (KNN). The Linear Regression model, using the selected 13 variables, achieved an impressive Root Mean Squared Error (RMSE) of 0.542 and an adjusted R-squared value of 0.9323. This means that the Linear Regression model was able to explain over 93% of the variance in the delivery lead time, demonstrating its exceptional predictive accuracy. In comparison, the KNN model, while also performing well, was outperformed by the Linear Regression model in terms of both predictive accuracy and interpretability.

Linear Regression Model -

```
> forecast::accuracy(predicted.best, logistics_validation$delivery_lead_time)
                    ME      RMSE       MAE         MPE    MAPE
Test set -0.01969293 0.542498 0.4224151 -0.6243155 13.705
```

K- Nearest Neighbors Model-

```
> forecast::accuracy(predicted_knn,logistics_test$delivery_lead_time)
                     ME      RMSE       MAE        MPE      MAPE
Test set -0.01462302 1.130595 0.9039881 -15.06652 29.51093
```

Model Evaluation

The final step in the modeling process was to evaluate the performance of the developed models. The team's analysis concluded that the Linear Regression model outperformed the KNN model across multiple evaluation metrics. The Linear Regression model not only provided better predictive accuracy, as evidenced by the lower RMSE, but it also offered greater interpretability. This means that the Linear Regression model was better able to explain the relationships between the predictor variables and the target variable (delivery lead time), making it a more valuable tool for logistics managers and decision-makers.

Conclusion

This project has successfully demonstrated the feasibility of predicting delivery lead time using machine learning techniques. Among the models evaluated, the Linear Regression model emerged as the most effective, achieving an RMSE of 0.542 and an adjusted R-squared value of 0.9323, outperforming the K-Nearest Neighbors (KNN) model.

The key highlights and contributions of this study are:

1. Comprehensive data preprocessing and feature engineering techniques were employed to ensure data quality and model robustness.
2. Variable selection methods, such as Best Subset Selection and Forward Selection, were used to identify the most significant factors influencing delivery lead time.
3. The Linear Regression model provided exceptional predictive accuracy and interpretability, making it a valuable tool for logistics managers and decision-makers.

Looking ahead, future work could explore the integration of advanced techniques like ensemble learning and the incorporation of external variables (e.g., weather, traffic patterns) to further enhance the predictive capabilities of the models. Additionally, the findings of this project can serve as a foundation for developing more sophisticated decision support systems in the logistics industry, ultimately leading to improved operational efficiency and customer satisfaction

References

-Project Dataset: Logistics Data
-Methodologies: Best Subset Selection, Forward Selection, Linear Regression, K-Nearest Neighbors
-Tools Used: R