Group 6 Project Report
Company - TKL Logistics

Aswathy Gopalakrishnan Deepa Krishnan
Jini Saroja Muhammad Malik
Punnoose Kozhuppakalam Thomas
Sneha Mathummal Parapurath Yonge Li
Reshmi Chandran Bhuvanachandran Nair Sobhanakumari

December 9, 2024

Contents

1 Introduction

TKL Logistics is a rapidly growing global logistics provider with a significant yearly shipping volume. The goal of this project is to enhance their freight booking efficiency through the use of predictive analysis, driven by AI. The solution focuses on generating well-informed decision choices that take factors like time, sustainability, and price into account. The report explains the various process in the development of the solution to help in the decision-making process of freight booking.

2 Business Understanding

The main focus of the project is the ETA predictability of data analysis with sustainability and price in mind. The main objectives of the project based on the proposal by the company are:

- Get better predictability of ETA

  The primary objective is to enhance the predictability of ETA using historical data. In logistics, accurate prediction is important for inventory management, customer satisfaction, and overall efficiency.

- Compare options in regards of price and CO_2 in real time

  The second objective involves comparing and analysing transportation based on price and CO_2 emissions. The goal focuses on improving sustainability and cost efficiency in logistics.

- Rank transporters in accordance to historic delivery performance

  The third objective to identify and rank transporters in terms of their historic delivery performance. This enables the freight booker to choose the best agents based on their performance metrics.

2.1 Data Access and Alignment with Freight Booking Process

One of the key questions during the Business Understanding phase is what data we have access to. That will help us set realistic goals and expectations for the analysis.

The link provided by TKL (dispatch.tkl.se) was explored for better understanding of the freight booking process. By reviewing the site, insights into how shipments are scheduled, how to select transporters, the factors influencing pricing and ETA are understood. This understanding will help us align the data analysis and AI models with the actual process of freight booking, ensuring our solutions are relevant and practical.

2.2 Key Observations

The business understanding phase has helped us to determine the key objectives that the improvement of ETA predictability, comparison of transportation options in terms of cost versus sustainability, and ranking transporters based on historical performance will be pursued. In this regard, the insights into the TKL system bring a basis for developing artificial intelligence models and data analytical tools that directly support these goals.

3 Data Understanding

The primary goal of data understanding is to obtain an extensive knowledge of the dataset in order to make sure that it is appropriate for analysis with the project goal. Five steps are included.

3.1 Data Collection

We received the dataset containing various features related to transportation logistics of TKL Logistics. This dataset served as the foundation for our analysis, and a critical initial step was to carefully examine its structure and contents to understand the provided parameters.

3.2 Data Description

The data provided in a structured format. The dataset includes key columns such as LogEntryIDs, Distances, Weights, OriginalETAs, TotalTransportCosts, EntryDates, MeansOfTransport, CO_2 emission, LoadingDates, UnloadingDates , and DeliveryDates etc.

3.3 Data Quality Verification

In our project, a critical initial step was to carefully examine the dataset to ensure they aligned with the project objectives. During this process, we identified several data quality issues that required attention.

1. Missing Values: Critical columns like distance and weight had missing data.
2. Invalid Data: Some entries lacked the original ETA and contained invalid dates.
3. Inconsistent Data: Issues regarding zero or negative transport costs were noted.

3.4 Data Exploration

We examined the TKL logistics dataset to understand its structure and analyzed the relationships between the various features using graphical visualization. These findings helped us identify potential data quality issues that needed to be addressed before proceeding with analysis. It was also identified that a few relevant columns, like weight, were missing.

3.5 Communication with the Business Team

To address our concerns about the data provided, we communicated our observations and questions to the business team, the following are a few among the questions asked:

- Should container deliveries have the same delivery days as ships?
- Are flights exclusively used for express delivery?
- Is the entry date a necessary parameter?

3.6 Data Improvement

The business team provided the updated and corrected data, addressing the issues and inconsistencies identified during our initial analysis. Missing values in critical fields, such as distance and weight were filled or rectified, ensuring completeness. Corrections were made to transport costs, including clarifying that zero values represented cancellations and negative values indicated returned items. Express deliveries via flights were confirmed. These improvements enhanced the overall accuracy, consistency and reliability of the dataset.

4 Data Preparation

Data Preparation phase is essential to ensure that the data is clean, and in the right format for effective modelling.

4.1 Data Cleaning and Transformation

During the data preparation phase, several issues within the date columns such as EntryDates, LoadingDates, UnloadingDates, OriginalETAs, and DeliveryDates were identified.

- Several rows contained incorrect values for years. For example; `‘2323-07-13’`. To solve this issue, string manipulation was used to replace the year with proper values.
- The dates were all initially in string format. For example: `‘2021-01-15T00:00:00’`. These were changed to standard date format.
- The OriginalETAs column presented dates in a different format (`’11/12/2024 00:00:00’`). This was also converted to a similar date format to match the other date columns.
- Following the discussions in company meeting, it was determined that the EntryDates column was not required for the analysis. Hence, this column was filtered out from the dataset.

After the data cleaning, transformation, and reduction of the dates, data integration was done to produce new columns. The difference between the loading and unloading dates were calculated to determine the transport days. While, the difference between unloading and delivery dates were used to get delivery days. The sum of these two metrics provides the overall delivery time. However, there were issues with some rows showing negative values as there were some discrepancies in the date entries. On analysis, it was found that the dates were interchanged, or wrongly entered. This led to the miscalculations. To solve this, the estimated days provide by the company based on the means of transport were used to update the incorrect values. The column was updated using these values wherever necessary.

4.2 Distance and CO_2 Data Restoration

Most of the rows in the Distances column were empty. To solve this, a systematic approach was followed to fill the missing values. The FromZipCode,ToZipCode, and MeansOfTransports columns were used to group the dataset. Based on this, the corresponding available distance values from the distance column we taken and used to replace the empty entries in the column.

Similar procedure was followed to fill the empty values in CO_2s column as well. The dataset was grouped in similar manner and the CO_2 values are replaced with existing ones. This process ensured that the added distance and CO_2 values are contextually relevant, thereby preserving the integrity of the set. Both the columns were in string format. These were typecast to numbers.

4.3 Formatting Transportation Costs Data

Transportation dates were originally provided in string format (`'19679,4800 SEK'`). To process this data, first the 'SEK' part was removed using string manipulation, and then the rest was converted to number format.

During the analysis, it was noticed that some of the transportation cost values are zero, or negative. After consulting with the company representative, it was clarified that the zero values indicated cancellations and the negative values were for returns. As these entries are not relevant for the prediction, these rows were removed as part of the data cleaning process.

4.4 Feature Engineering for ETA Prediction

With the data cleaned and formatted, the next step was to select the appropriate columns for prediction. A linear correlation analysis was done to examine the relationship between the variables and the ETA column. To and From Zipcodes, CO_2s, Total Transport Costs, Number of Pieces, and Weights showed a linear correlation. However, the distance and agent columns did not exhibit a clear relationship. This could be because the relationship is non-linear.

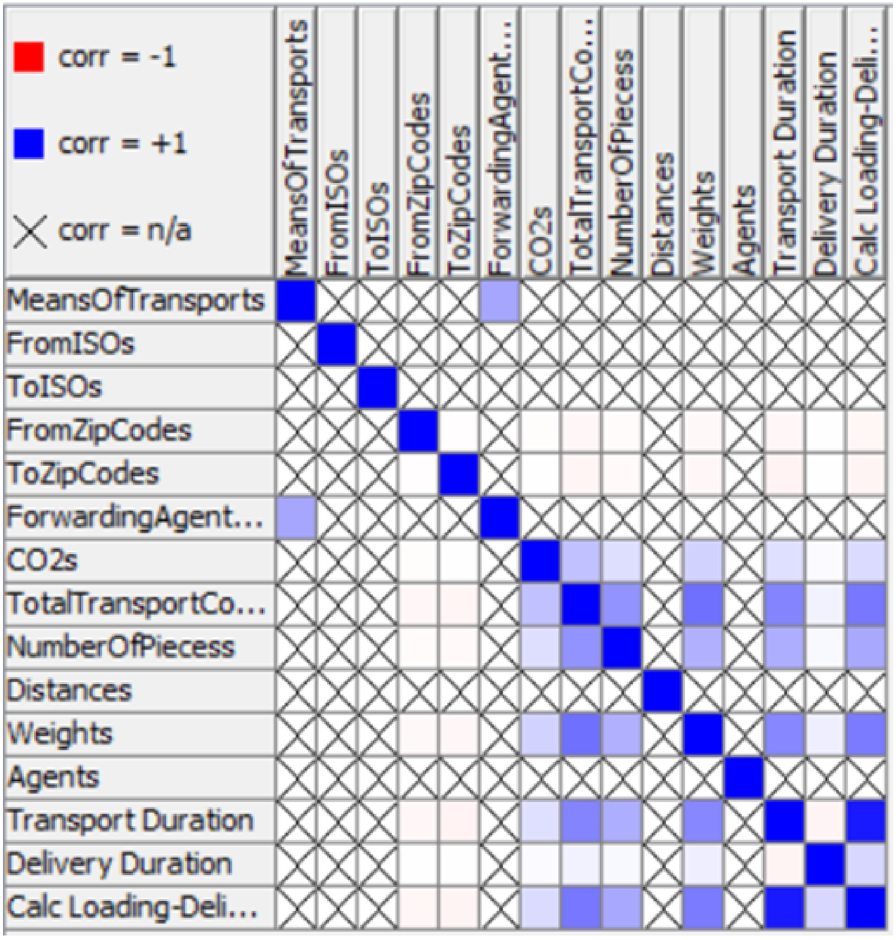

Figure 1: Correlation Analysis to understand relevant columns

To analyse this further, graphs were plotted to visualize the relationships. It was clear from the graphs that both affect the ETA. When there is an increase in the distance, there is an increase in the ETA too. The clustering was done using the means of transport. It showed that the means of transport varied the ETA even though the distance was the same. For example, the same distance covered by air transport and container transport provided two different ETAs. With these processes, the required columns were successfully identified for the prediction.

Figure 2: Plot 1 : Distance Vs ETA. Plot 2 : Agents Vs ETA

4.5 Final Data Formatting

In the final stage of data preparation, unnecessary columns were filtered out. Any missing values that remained were handled using the missing values node. Categorical data was converted into a numerical format. Min-max normalization was applied to scale the data within a range of 1 to 10, to improve the comparability of different features.

Finally, any negative entries in Distance, Transport Cost, and CO_2 emissions (if any) columns were removed. The final data is then passed into the regression models.

Rows: 75799 | Columns: 12

#	RowID	Means... Number (d...	FromISOs Number (d...	ToISOs Numbe...	FromZipC... Number (dou...	ToZi... Numb...	CO2s Number (dou...	TotalTran... Number (dou...	NumberO... Number (dou...	Distances Number (dou...	Weights Number (dou...	Agents Number (dou...	Calc Loa... Number (dou
1	Row0	1	1	1	1	1	1.007	1.268	1.033	6.811	1.032	1	5.16
2	Row1	1.562	1.15	1	1.004	1	1.016	1.378	1.38	5.398	1.088	1.026	6.824
3	Row2	1.562	1	1	1.002	1	1.016	1.386	3.404	7.529	1.646	1.053	5.311
4	Row3	2.125	1.15	1	1.004	1	1.032	1.559	1.015	3.623	1.403	1.026	6.294
5	Row4	2.688	1.3	1.083	1	1	1.025	1.326	1.045	6.638	1.273	1.053	4.933
6	Row5	2.125	1	1	1.001	1	1.03	1.476	1.008	7.362	1.66	1.079	5.613
7	Row6	2.688	1.3	1	1	1	1.026	1.798	1.026	6.469	1.307	1.079	4.782
8	Row7	2.688	1.3	1	1	1	1.026	1.798	1.029	6.469	1.295	1.079	4.782
9	Row8	2.688	1	1	1.002	1	1.031	1.816	1.66	6.934	1.283	1.106	3.723
10	Row9	2.688	1	1	1.002	1	1.03	1.752	1.476	7.463	1.338	1.106	4.252
11	Row...	2.125	1	1	1.001	1	1.03	1.524	1.194	6.862	1.188	1.106	5.387
12	Row...	2.125	1	1	1.001	1	1.03	1.466	1.261	6.862	1.189	1.106	5.387
13	Row...	1	1	1	1.002	1	1.002	1.08	1.001	7.315	1.033	1	5.387
14	Row...	1	1	1	1	1	1.003	1.077	1.001	6.811	1.037	1	5.311
15	Row...	1.562	1	1	1.002	1	1.015	1.501	1.018	7.194	1.7	1.026	3.95
16	Row...	1	1	1	1.002	1	1.002	1.14	1.003	6.811	1.053	1	3.723
17	Row...	2.688	1	1	1.001	1	1.03	1.44	1.311	6.934	1.206	1.106	5.916
18	Row...	2.688	1	1	1.001	1	1.03	1.44	1.296	6.934	1.199	1.106	5.916
19	Row...	2.688	1	1	1.001	1	1.03	1.44	1.296	6.934	1.199	1.106	5.916
20	Row...	2.688	1	1	1.001	1	1.03	1.493	1.426	6.934	1.824	1.106	5.992
21	Row...	1.562	1	1	1.002	1	1.016	1.323	1.008	7.529	1.187	1.053	4.933
22	Row...	2.688	1	1	1.001	1	1.03	1.493	1.402	6.934	1.464	1.106	5.992

Figure 3: Dataset after data preparation

5 Modelling

The key phase of the CRISP-DM methodology is modelling, as it involves developing and refining machine learning models that can reliably forecast ETA prediction for freight booking. To meet business requirements, we have chosen a variety of regression methodologies.

5.1 Modelling technique and modelling assumptions

For freight ETA prediction using stratified sampling based on means of transport, we have considered Random Forest, Gradient Boosted Trees and Simple Regression Tree regression models. To improve model performance, we have stratified the sample by means of transport. We have made a few assumptions about the data to ensure that no missing values are allowed for the existing ETA and instead of predicting the exact date of arrival we are calculating days. We have also removed outliers from the dataset which was mainly due to invalid date entries.

Random Forest Regression : This learning technique can handle more complex and non-linear relationships in the booking data, and hence it is highly successful for ETA prediction. It provides insights into feature selection to understand key factors influencing ETA prediction. Additionally, it can also handle missing values, which is beneficial in real-time freight booking.

Gradient Boosted Trees : It is another powerful ensemble method which works well for ETA prediction considering both historical and real-time data. It outperforms more straightforward models like decision trees and simple regression tree in long-term forecast.

Simple Regression Tree : It is a sophisticated modelling algorithm which is easier to interpret compared to other ensemble methods. It can also capture non-linear relationships in the data. It is less accurate than Random Forest Regression and Gradient Boosted Trees for predictive tasks with complex logistics situations.

5.2 Test design

To evaluate the performance of the chosen modelling techniques for ETA prediction, we trained and tested models under similar conditions. All models were trained using 10-fold cross validation on the same processed data. The assessment criteria considered were R squared, mean absolute error and Root mean square error.

For ETA prediction we adapted the evaluation metrics as follows:

1. R Squared - To assess how well the models explain the variance in ETA predictions
2. Root Mean Square Error - Measures the standard deviation of the prediction errors, providing insight into the model's precision.
3. Mean Absolute Error – To measure the average absolute difference between predicted and actual ETA.

The model with highest R squared and the least Root Mean Square Error and Mean Absolute Error would be selected as the final solution for ETA prediction. This approach ensures that we choose the most accurate and reliable model for estimating the arrival time of freight.

5.3 Build model

We have used the following features for training the model:

- Means of transport
- From ISO

- To ISO
- From Zip code
- To Zip code
- CO_2 emission
- Total transport cost
- Number of pieces
- Weight
- Distance
- Agent

We have removed unwanted features from training the model to reduce the noise.

5.4 Parameter Settings

For freight ETA prediction models, key parameters and their chosen values are:

Random Forest

- Number of models: 100
- Limit number of levels: 20
- Minimum node size: default (5)

Here the number of model settings balance model complexity with computational efficiency.

Gradient Boosted Trees

- Number of models: 100
- Limit number of levels: default (4)
- Learning rate: 0.1

lower learning rate of 0.1 allows for a more robust model by reducing the impact on each tree.

Simple Regression

- Missing value handling: XGBoost

- Limit number of levels: default (10)
- Minimum split node size: default (1)
- Minimum node size: default (5)

A depth level of 10 allows for complex relationships to be captured without overfitting whereas minimum node size helps model to make generalized predictions.

5.5 Assess model

Based on our assessment, the Random Forest regression model demonstrated superior performance for ETA prediction with least Root Mean Squared Error and Mean Absolute Error, whereas R Squared value was highest as shown below.

RowID	ETA Prediction - Gradient Boosted Tree Number (double)	ETA Prediction - Random Forest Number (double)	ETA Prediction - Simple Regression Number (double)
R^2	0.912	0.928	0.866
mean absolute error	0.3	0.27	0.348
mean squared error	0.301	0.248	0.458
root mean squared error	0.549	0.498	0.677
mean signed difference	-0.031	-0.001	0.008
mean absolute percentage error	0.111	0.1	0.131
adjusted R^2	0.912	0.928	0.866

Figure 4: Numeric Scorer for Gradient Boosted Tree, Random Forest and Simple Regression models

The tree depth of 20 and minimum node size of 1 allows the model to capture complex patterns in the data, which is important for predicting ETA accurately based on various transport modes. The depth limit also helps in reducing overfitting the model. We tried changing the tree depth to 50 and minimum node size to 3, but it didn't make any impact on the model's performance.

We have experimented with the number of models beyond 100 (tried 200 and 500) but resulted only in marginal improvement in terms of evaluation metrics, while this change significantly increased computational time. So, we kept the number of models parameter value as 100 which provides a good balance between model accuracy and efficiency. Additionally, we attempted changing the learning rate of gradient boosted tree from 0.1 to 0.05 and increased number of models to 200, but it had no effect.

6 Evaluation

6.1 Predictability of ETA

The workflow developed in KNIME Analytics Platform aimed to identify the best model for predicting total duration or ETA in the dataset. Three different regression models were implemented and evaluated – Random Forest, Simple Regression Tree and Gradient Boosted Trees. Each model's performance was compared using the evaluation metrics such as accuracy and error. Model Comparison:

1. **Random Forest**

   This model delivered the highest accuracy with an R^2 value of 0.928. It also had the lowest error rates across the parameters, including mean absolute error (0.248) and root mean squared error (0.498). This consistency highlights its robustness and sustainability for the task.

2. **Simple Regression**

   The simple tree model had the lowest accuracy with an R^2 value of 0.866. It also had the highest error rates such as a mean squared error of 0.458 and root mean squared error of 0.677. These results indicate its limited capacity in capturing complex patterns compared to other models. Inversely, low scores for simpler models like the Simple Regression Tree indicate the ETA prediction problem to be of a complex nature.

3. **Gradient Boosted Trees**

   This model achieved an R^2 of 0.912, slightly lower than Random Forest. While the performance was still competitive, its errors were slightly higher than the top-performing models.

6.2 Key Takeaways

- The Random Forest model is recommended for its superior accuracy and lower error rates.
- Gradient Boosted Trees provide strong alternatives with slightly lower performance but are still suitable for many applications depending on priorities.
- The Simple Regression Tree model was the weakest performer and may not be suitable for complex datasets like the TKL Logistics dataset.

Overall, this evaluation demonstrates the value of comparing multiple models to select the most effective solution for predictive analytics tasks in logistics data.

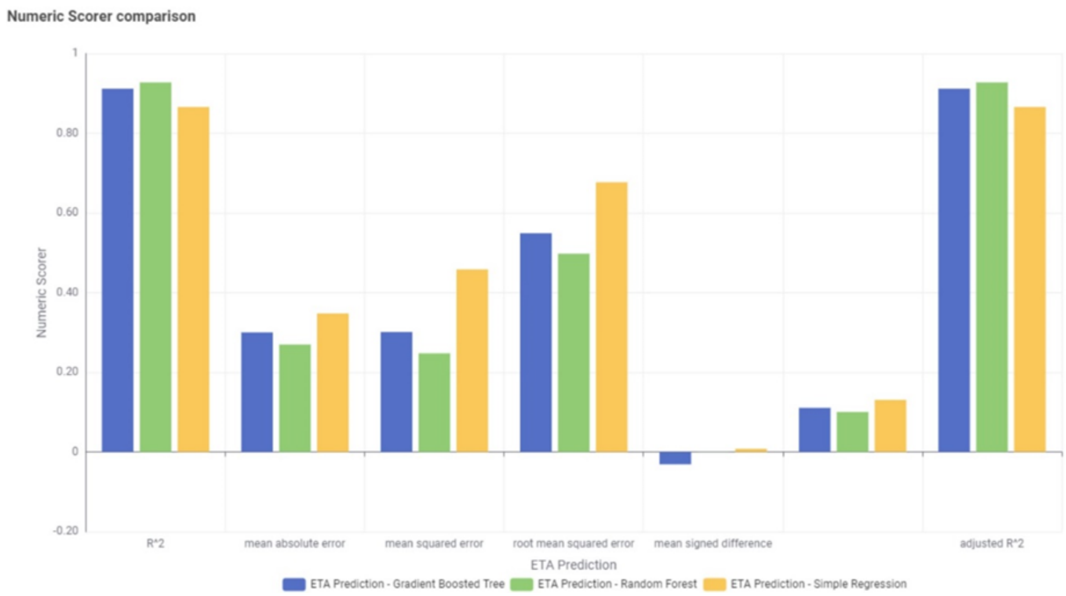

Figure 5: Comparison of scores for all 3 models

Figure 6: Predicted ETA Vs Actual ETA, Pie chart to show the relevant features and their contribution in the prediction

6.3 Evaluation of factors affecting ETA prediction

Feature importance in Random Forest model helps us to identify the most crucial factors affecting ETA prediction. From the pie chart above, it is clearly seen that means of transport, distance, agents, weights, and transportation costs constitutes major deciding factor. We can do fair prediction if transport mode and the distance is known. By identifying the crucial features, prediction models can be optimized for better performance and reduce the noise caused by non-relevant factors.

6.4 Evaluation of factors affecting price and CO_2 emissions

Clustering was done based on means of transport to find how price and CO_2 emissions were affected based on other factors.

The CO_2 emissions increase with increase in distance and weight. It is also affected by the means of transport. For instance, it is the highest for air transport even for shorter distances. It is the lowest for truck and comparatively lower for containers too. Hence, a conclusion was made to filter the means of transport based on the distance and weight.

The price is also affected by the same factors; means of transport, distance and weight. In addition to this, the ETA is also a factor in deciding the price. For express delivery, the air transport is preferred. This is higher in price and CO_2 emissions. However, consider the ETA, no other options can be selected. The truck shows lower CO_2 emissions and transport cost. Also, it can carry more weight than air transport. Containers take a longer ETA with low CO_2 emissions and comparatively less transport cost compared to air. Hence, it was understood that all these factors are interdependent and combined strategy must be followed to develop the selection process.

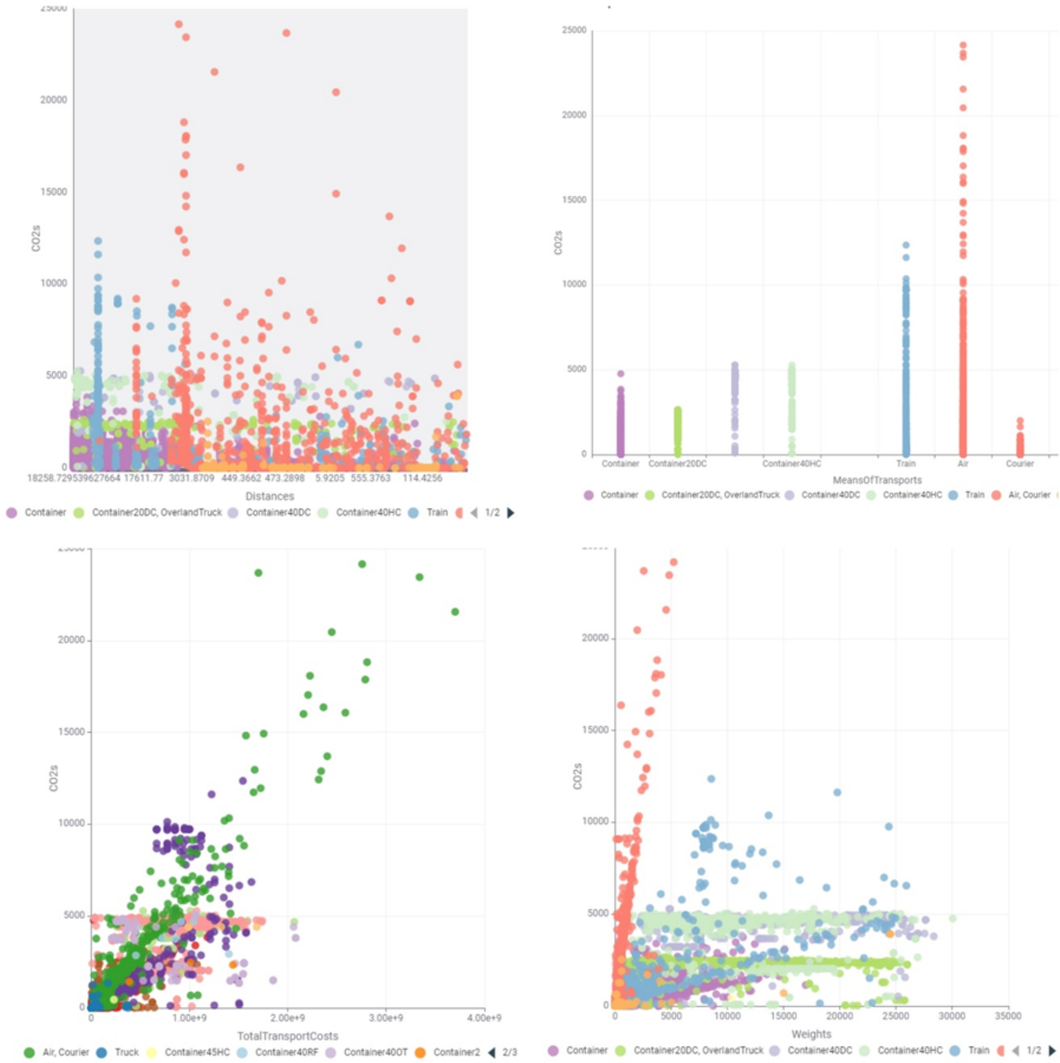

Figure 7: Clustering based on means of transport to analyze CO_2 emissions during transport

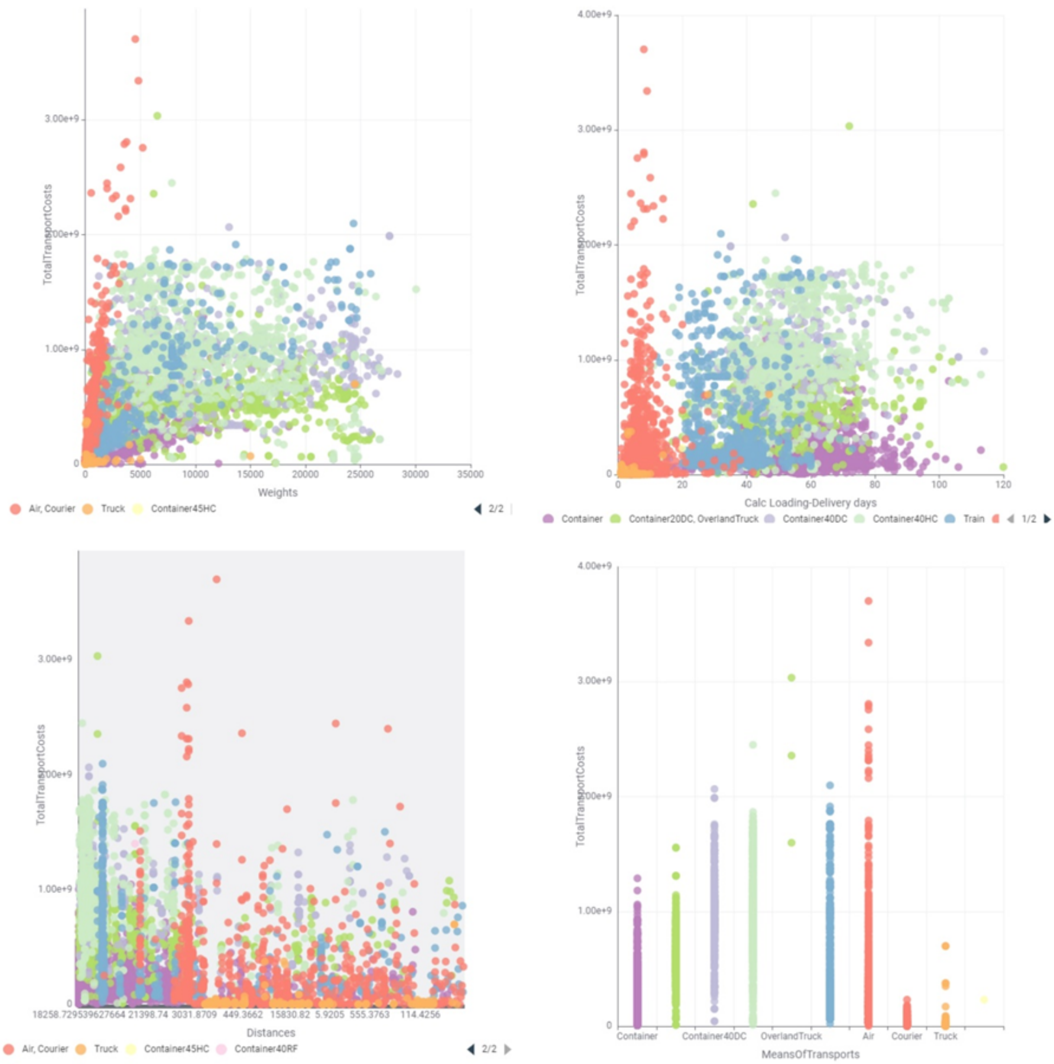

Figure 8: Clustering based on means of transport to analyze transport cost

7 Deployment

This is the final phase of the CRISP-DM methodology. At this stage, we have finalized the model which we will be using for deployment, in our case it is Random Forest Regression model. Deploying and making it accessible to external customers can be accomplished in several ways. We considered deploying the model as a service, but we lacked access to KNIME Business Hub. To address this limitation, we have created a process that uses the Model Writer node to save our model and the Model Reader node to read it to forecast the ETA in real-time scenarios. Missing values and incorrect data will be handled during preprocessing stage for new unknown data.

8 Suggestions

8.1 Experimental Setup for Predicting Estimated ETA

In this experimental setup, our objective was to predict the estimated time of arrival (ETA) for packages or freight shipments. We evaluated various machine learning models for their performance in making accurate predictions. Among the models tested, the Random Forest model demonstrated the best performance based on the parameters we analysed.

While our model showed promise in making better predictions, it is essential to account for real-time variables that can influence ETA accuracy. Factors such as weather conditions and national events or issues in both the origin and destination countries play a significant role. These factors can disrupt transportation schedules and delivery timelines, thereby impacting the reliability of ETA predictions. Currently, the absence of such real-time data in our dataset limits the model's ability to account for these influences.

8.2 Limitations in Dataset and Agent Ranking

In the process, we tried ranking the agents based on the location, ETA, distance, CO_2 emissions, and means of transport. However, it didn't seem fair to rank the agents based these factors. The type of services they provide, how they handle the shipments, the number of shipments they handle on a daily basis, the resources they have, etc are all factors that affect their ranking. Due to insufficient data, ranking delivery agents based on their performance is not feasible. Adequate data about agent-specific factors, such as delivery efficiency or service disruptions, would be crucial for implementing a ranking system. Without this, the model's insights remain generalized.

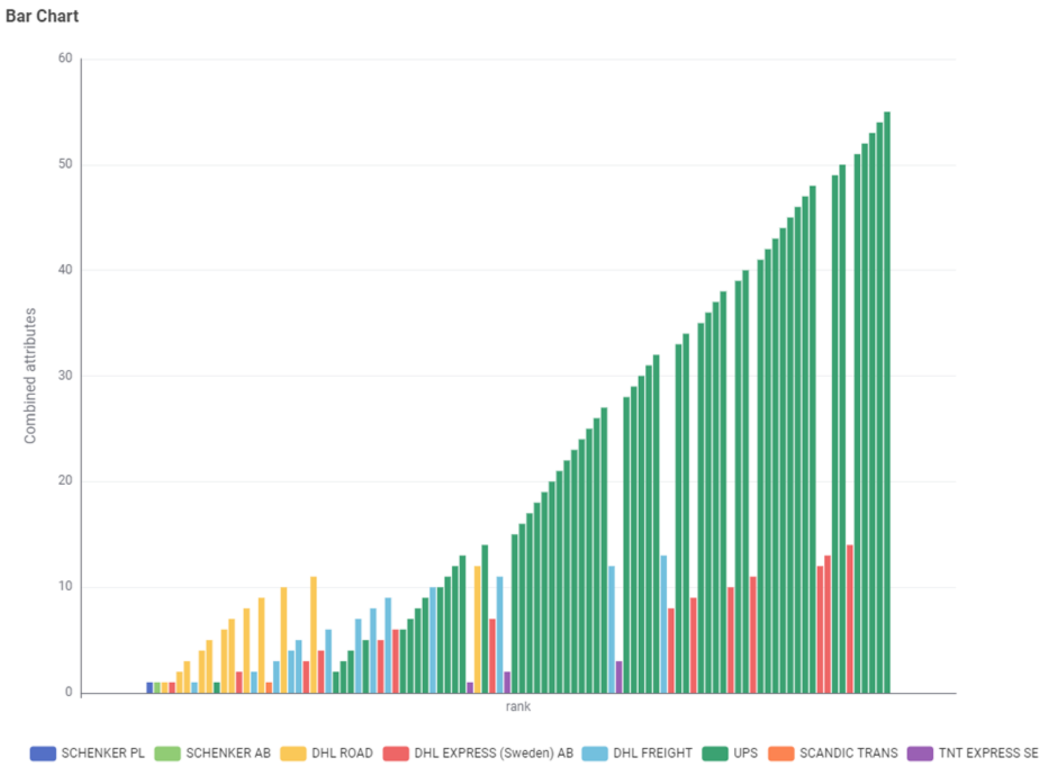

Figure 9: Ranking Agents based on relevant attributes

8.3 Suggestions for Dataset Improvement

To enhance the dataset's quality and ensure more robust model predictions, we recommend the following:

- Include Cancellation and Return Data:

  The dataset should include information on whether a package or freight was cancelled or returned. This addition would provide valuable insights into logistical bottlenecks and help evaluate delivery success rates more comprehensively.

- Validate Date Entries:

  Dates must be validated at the data entry stage to avoid incorrect or outlier values. This step ensures data integrity and prevents anomalies that can distort the model's predictions.

- Enhance Application Usability:

  The interface of the freight booking application can be improved with the following features:

- Input Prioritization: Arrange tabs so that to and from zip codes (generates distance), expected delivery date for customer, and type and specification of the goods, are entered first.

- Filtered Transport Options:Based on the input, filter transport options by type, prioritizing those with lower CO_2 emissions and price.

- Promoting Sustainability:Display options with lower CO_2 emissions prominently, enabling users to make environmentally conscious decisions. This feature can contribute significantly to reducing CO_2 emissions over time.

9 Conclusion

By addressing the outlined real-time factors and improving the dataset, we can enhance the accuracy of ETA predictions and implement a reliable agent-ranking system. Additionally, optimizing the freight booking platform with sustainability-focused features will not only improve user experience but also align with broader environmental goals.