TABLE OF CONTENTS

1. Executive Summary

J&T Express is a global logistics service provider with leading express delivery businesses in Southeast Asia and China, the largest and fastest-growing markets in the world. In Singapore, as a one-stop e-commerce specialist, J&T Express offers three core e-commerce solutions catered to online-businesses in the market, namely - last-mile delivery, fulfillment and international delivery.

J&T has been experiencing feedback from customers regarding delayed deliveries and parcels languishing in J&T's hub for prolonged durations. With the e-commerce sector experiencing rapid growth and becoming increasingly integral to the global economy. It is urgent and significant for J&T Express to address these shortcomings to stay competitive in the delivery sector. The root cause of these problems stems from an inefficient route planning system and failure to anticipate the surge in e-commerce demand.

We aim to tackle the operational challenges faced by J&T through the use of analytics. The development of predictive models would help J&T reduce delivery time of customer parcels from the delivery hub to their locations by forecasting seasons of surge and dip in demands for ecommerce goods, and optimizing route planning.

J&T Analytics is a software that integrates a demand forecasting model with a route optimisation model. The software was created in a 3-step process. Firstly, we utilized a dataset provided by a Brazilian department store, Olist, to perform data exploration and analysis to identify variables that could affect the time taken for ecommerce deliveries to be completed.

Secondly, we sought to find a model that we could use for demand forecasting. We tested different models which include, Random Forest, Neural Network and XGBoost, and determined the optimal model by comparing R-squared, Root Mean Square Error (RMSE) and Mean Absolute Error (MAE). We then compared 2 clustering models, parcel clustering and location clustering, to evaluate which type of clustering model is a better fit for route optimization analytics.

Finally, we integrated the 2 models into a single shared software, where J&T's staff have access to real-time analytics and insights. J&T Analytics will provide insights on the predicted size of orders for each month which allows J&T Express to react quickly and anticipate changes in demands for better and more efficient allocation of resources. The software will also be linked to an app that will provide the relevant information needed for drivers, to ensure that they arrive at the destination as soon as possible.

J&T Analytics is sustainable and perpetually relevant as it is constantly being updated and refreshed with the latest trends and findings from the real-time data it collects. We believe that this software will be able to revolutionize J&T's strategies to improve its services and operations, which will ultimately bolster business performance and alleviate profit margins, hence enhancing the company's competitive edge in the market.

2. Introduction

2.1. Business Problem

J&T Express has been grappling with a significant challenge that has led to widespread customer dissatisfaction: a notable delay in parcel delivery times. This problem has been particularly acute from June 2021 to February 2022, during which customers frequently reported their parcels languishing in J&T's hub for prolonged durations without any communication or resolutions from the company. This issue primarily arises from an underestimation of the burgeoning e-commerce demand , a challenge exacerbated by the COVID-19 pandemic. The company's reliance on manual processes for route planning coupled with limited insights into future demand patterns, has critically hampered the company's ability to meet delivery times efficiently.

In response to this pressing issue, J&T Express is poised to undertake a strategic pivot aimed at fundamentally enhancing operational efficiency and customer satisfaction. By leveraging on advanced predictive modeling techniques, including demand forecasting and optimizing route planning strategies, J&T Express intends to proactively address the root cause of the delivery delays. This initiative is not just about responding to the current crisis but to reimagine how J&T operates in the fast-evolving e-commerce landscape to remain competitive. Through the adoption of these technological solutions, J&T Express is committed to transforming its delivery infrastructure, ensuring that it can meet customer expectations with agility and precision, thereby fortifying its position in the competitive logistics market.

2.2. Significance of Business Problem

Singapore, being one of the largest e-commerce markets in Southeast Asia, it is unsurprising that many businesses are keen to catch this very promising and lucrative market opportunity. The vibrant market presents not only a tremendous opportunity but also a complex array of challenges that test the resilience and agility of companies like J&T Singapore. In such a competitive environment, the efficiency of logistics and excellence in customer service are not just operational goals but critical business imperatives. Minor logistical setbacks, such as suboptimal route planning, can cascade into significant issues, including missed deliveries, dissatisfied customers, and tangible impacts on revenue and beans reputation.

The primary concern for delivery companies is the delay in deliveries, a persistent issue that undermines the reliability of logistics services. A spokesperson for J&T Singapore emphasized that "delays and other delivery issues remain a key challenge for all logistics players." Many unexpected factors such as weather conditions and traffic jams, together with a surge in parcel volumes during sale campaigns and festive seasons, exacerbate this challenge.

Moreover, an analysis into Feature Importances showed that 'price' and 'delivery status' are some of the main drivers for review scores. E-commerce platforms in this case J&T can use this knowledge to enhance customer satisfaction since it points out how pricing strategies matter and the importance of delivery process efficiency.

	feature	importance
1	price	0.554627
5	x0_delivered	0.172593
0	order_item_id	0.061327
59	x1_housewares	0.013071
76	x1_sports_leisure	0.012193
53	x1_health_beauty	0.010871
17	x1_bed_bath_table	0.008708
15	x1_auto	0.008692
52	x1_garden_tools	0.007812
49	x1_furniture_decor	0.007681

Figure 1: Feature Importance Scores

The repercussions of J&T Singapore's struggles with delayed deliveries and subpar customer service as mentioned in the business problem extend beyond immediate customer dissatisfaction. It will affect the company's reputation and reliability, which could cost the firm its existing and potential partnerships with ecommerce merchants and platforms. It will also undermine its competitive standing in strategic markets such as Southeast Asia and China, as competitors such as PickUpp, which offer insured packages, and NinjaVan, which offer extremely competitive starting prices from $2.50/kg, continue to outshine and rob customers and opportunities from J&T express.

Addressing these complex logistical challenges necessitates deep dive into data analytics to uncover underlying issues and craft innovative solutions. Employing machine learning models for predictive demand forecasting and route optimisation emerges as a strategic necessity. Such technologies offer a dual advantage: foresight to avoid potential disruptions and enhanced operational agility. In a marketplace where differentiation is a key to success, leveraging advanced analytics and machine learning can propel a company to the forefront, ensuring it not only meets but exceeds customer expectations in a highly competitive sector.

2.3. Current Framework and Limitations

The current framework within J&T Express Singapore for handling deliveries, as described, seems to heavily rely on traditional and possibly outdated methodologies. This framework is likely to include manual route planning and a lack of proper real-time data analytics for demand forecasting. Let's now discuss the limitations of this current framework in detail:

2.3.1. Manual Route Planning:

- Inefficient Resource Management: There is risk for over allocating manpower due to the absence of proper route optimisation. This would lead to unnecessary costs and potentially having underutilized personnel.
- Inefficiency in adapting to real-time conditions: Manually planning routes is time consuming and often less efficient compared to automated systems that can dynamically adjust to real-time conditions such as traffic, weather, accidents and sudden changes in delivery schedules.
- Human Error: There's the obvious risk of human error, including suboptimal route decisions that can lead to increased delivery times and fuel consumption.
- Issues with Scalability: As parcel volumes grow, especially during peak periods/seasons, manual route planning becomes increasingly unmanageable and time consuming and cannot scale effectively to meet the demand as more and more manpower would be required.

2.3.2. Lack of Real-time Demand Forecasting:

- Inaccurate Predictions: Without advance demand forecasting techniques, J&T might not accurately predict spikes in parcel volumes, leading to underpreparedness during peak periods.
- Resource Misallocation: Without a clear forecast of demand, resources (like delivery personnel and vehicles) may be misallocated, with some areas having more resources than needed and others experiencing shortages.
- Reactive Instead of Proactive: The absence of predictive analytics means the company is always reacting to situations rather than being proactive. This reactive approach can lead to delays and dissatisfaction from customers.

2.3.3. Excessive hiring of staff

- J&T Express hires both full- and part-time drivers as well as warehouse staff, to help manage surges in demand which helps to increase operational capacity and the quality of our delivery services. They have also been ramping up training for employees and putting in place more protocols, from pickup to sorting and delivery.
- Limitations: hiring excessive staff may not be efficient due to lower cost savings and lower productivity of workers during non-peak season

2.3.4. Revamp of internal processes

- Apart from purchasing extra warehouse facilities and vehicles, they also leverage technology by upgrading automated sorting machines and transport management systems, ensuring greater efficiency and accuracy. On the operations side, J&T aims to increase productivity and efficiency in their sorting hubs by relooking their workflow in terms of operations SOPs, warehouse layout, manpower deployment and sorting strategy.
- Limitation: Upgrading of machines and revision of operational procedures only help to enhance the efficiency of the internal processes. It does not solve the issue of the backlog parcels stuck at the sorting and distribution hub.
- JMS is J&T's smart transport management system that offers end-to-end visibility into the entire logistics chain, ranging from order placement to shipment execution and fleet management (Singapore Business Review, 2024). JMS' in-built data analytics helps us further optimize day-to-day processes and ensures the reliability of delivery services. (Martech Asia, 2023)

In conclusion, J&T Singapore currently faces operational challenges that hinder our ability to fully capitalize on the market opportunities. These challenges include manual route planning, poor customer service practices, absence of real-time demand forecasting, inefficiencies due to potential overstaffing, and the requirement for a more complete solution to deal with internal operational procedures. While efforts to enhance productivity through workforce expansion, the integration of JMS (J&T Management System), and technological upgrades have yielded improvements, we continue to grapple with scalability issues, human error, and reactive problem-solving.

To transcend these limitations and achieve a paradigm of sustained customer satisfaction and operational superiority, a strategic overhaul is imperative. J&T must embrace a holistic strategy that leverages the full spectrum of advanced analytics, embraces automation, and refines internal processes for maximum efficiency. This strategic realignment will not only empower us to meet the current demands with enhanced agility but also equip us to proactively navigate and thrive in the rapidly evolving landscape of the logistics industry.

By committing to this comprehensive strategy, J&T Express Singapore will not just address existing operational challenges but will set a new benchmark for innovation and service excellence in the logistics sector. This is our path to not only sustaining but elevating our competitive advantage, ensuring that we remain at the forefront of delivering exceptional value to our customers and stakeholders in an ever-changing global marketplace.

2.4. Our Approach

In response to the operational challenges identified above, ranging from manual route planning inefficiencies to suboptimal staffing strategies, our approach is centered around a transformative strategy that harnesses the power of data analytics. Our strategic approach integrates two pivotal solutions: demand forecasting and route optimisation through clustering, which will be integrated into one major solution. This dual-faceted strategy is designed to tackle our current operational bottlenecks head-on.

2.4.1. Demand Forecasting

Our first strategic initiative involves the implementation of demand forecasting models. By leveraging historical data and predictive analytics, we aim to forecast monthly demand. We will be testing machine learning models including Neural Network, Random Forest and XGBoost. Each model brings unique strengths in handling complex datasets and revealing insightful patterns, allowing us to determine the most effective approach by comparing the models. This precision in forecasting will enable us to determine the optimal number of drivers required, effectively aligning our staffing levels with actual demand. This approach not only ensures that we are prepared to meet the fluctuating needs of our market but also significantly reduces the risk of overstaffing, thereby optimizing our operational costs.

2.4.2. Route Optimisation

Leveraging the precise demand forecasts, our second strategic initiative targets the efficiency of our delivery routes and process through the application of clustering algorithms. This sophisticated combination allows us to refine our understanding of delivery dynamics. Clustering algorithms are employed to categorize deliveries into optimally groped clusters based on the parcel volumes as well as the geographical proximity and predicted delivery timeframes. Clustering helps in organizing deliveries in a manner that maximizes the efficiency of our routes and parcel loading, reducing unnecessary travel, and thereby significantly cutting down on delivery times and fuel consumption.

2.4.3. Integration Approach

Finally, the integration of demand forecasting and route optimisation represents a holistic response to the challenges previously outlined. This approach addresses the core issues of manual route planning, inefficient resource allocation and the lack of proactive operational strategies. By leveraging a shared

analytics platform, predictive insights directly inform and optimize delivery routes, ensuring adaptability and efficiency in real-time demand changes.

In conclusion, our strategic approach is not just about overcoming current limitations; it's about propelling J&T Express Singapore towards a future where we lead through innovation, efficiency, and unwavering commitment to customer service. With demand forecasting ensuring a leaner, more agile operation and route optimisation enhancing customer satisfaction through faster delivery times. We are poised to not only meet the dynamic demands of the logistics sector but to redefine them, ensuring our leadership position in the marketplace.

3. Data Exploration And Analysis

3.1. Data Exploration

This dataset comprises ecommerce orders of Olist, the largest department store in Brazil. Olist connects small businesses across Brazil to channels where merchants are able to sell their products through the Olist Store and ship them directly to the customers through Olist logistics partners. The dataset covers over 100,000 Olist orders from 2016 to 2018 and consists of customer, seller and product information.

Trend of orders

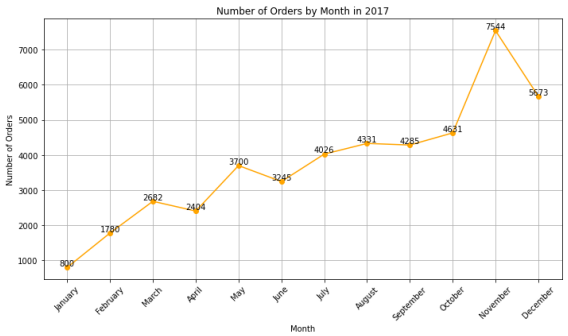


Figure 2: Line graph of order quantity in 2017

In 2017, there is an increasing trend in the number of orders through the year. Number of orders increased from 800 in January 2017 to 5673 in December 2017. There was a spike in the number of orders from 4631 in October to 7544 in November which could have been attributed to preparations for the festive season such as Christmas and winter holidays. The peak number of orders in November could have also been contributed by the Black Friday sale in November that offers product discounts so that customers can shop for holiday gifts.

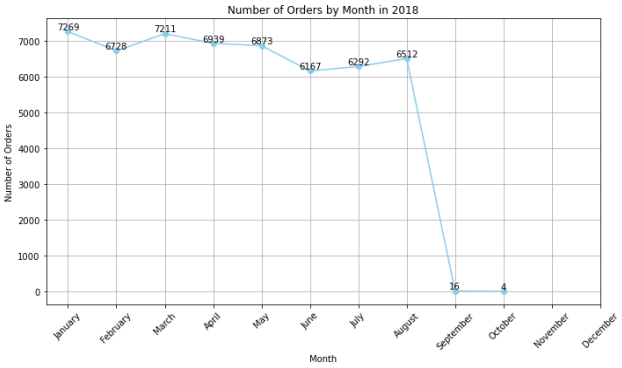


Figure 3: Line graph of order quantity in 2018

In 2018, the number of orders remained in a relatively high range from about 6000 to 7300 orders per month. Compared to 2017, the drastic increase in orders revealed that the ecommerce firm, Olist, is growing in popularity and also signifies the prevalence of increasing ecommerce usage in Brazil. January and March are months that have been identified to have a higher than average number of orders in 2018. The higher numbers could have also been attributed to the changes in seasonal demands due to different festivities and sales. For instance, the higher number of orders in January could be due to New Year sales and in March, the celebration of Brazilian Carnival that usually takes place between the end of February to the beginning of March.

Days to delivery

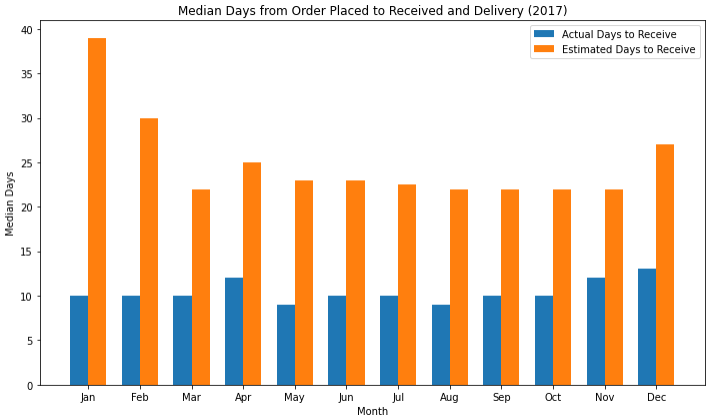


Figure 4: Bar chart of actual and estimated days to receive parcel

By analyzing the monthly median number of days customers took to receive their orders compared to the estimated number of days in 2017, there is a large discrepancy between both values. This shows that Olist's current predictive model may not be reliable and accurate in predicting the estimated date of arrival of orders, which calls for a revision of their existing model for possible improvements and enhancements.

Location of orders

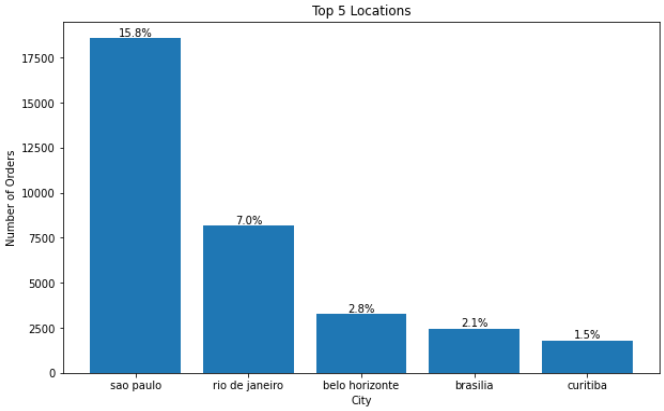


Figure 5: Bar chart of top 5 locations by order quantity

Based on the bar chart, 15.8% of total orders are from Sao Paulo, which is Brazil's financial hub and most populated city of about 11 million inhabitants as of 2022 (Statista, n.d.).

Insights from Data Exploration

1. **Ecommerce Growth and Seasonality**
   The data reveals a clear trend of growing ecommerce popularity in Brazil, as evidenced by the significant increase in the number of orders from 2016 to 2018. The seasonality is also highlighted as there are spikes in the average orders during months with festive events.

2. **Delivery Performance Issues**
   The discrepancy between estimated and actual delivery times points to potential issues in supply chain management and predictive modeling accuracy.

3. **Geographical Distribution of Orders**
   With Sao Paulo accounting for a significant portion of orders, it indicates a concentration of e-commerce activity in certain areas. This insight could guide logistics and distribution strategies, focusing on optimizing operations in high-demand regions.

3.2. Data Cleaning

The Olist dataset consisted of 8 separate csv files. However, we wanted to consolidate it into a large file with the relevant information needed for our models. Olist_order_payments was dropped as payment methods were not required for both models.

File	Columns dropped	Reason
olist_orders	Orders_approved_at Order_delievered_carrier_date Order_purchase_timestamp	The goal is to analyze delivery performance based on actual delivery dates. Only the final delivery date is relevant.

File	Columns dropped	Reason
olist_order_items	Seller_id Shipping_limit_date freight_value	The analysis is focused on the customer rather than the seller. Shipping information and freight value as we focus is not on shipping but rather delivery on land.
olist_order_reviews	Review_comment_title Review_comment_message Review_comment_date review_comment_timestamp	The sentiment analysis of reviews is not a part of the predictive model. The model is more focused on quantitative factors rather than qualitative feedback.
olist_products	Product_name_length Product_description_length	Analysis of product name and description is not relevant to the predictive models.
olist_customers	Customer_unique_id	Customer identification is not required for the analysis.

3.2.1. Geolocation

```
geolocation = pd.read_csv('/Users/Megan/Desktop/Y3S2/BC2407 Course Materials/project/data/olist_geolocation_dataset.csv')
geolocation = geolocation.groupby(['geolocation_city','geolocation_state','geolocation_zip_code_prefix']).last()
new_data = pd.read_csv('/Users/Megan/Desktop/Y3S2/BC2407 Course Materials/project/data/orders_forecasting.csv')
```

Figure 6: cleaning of Geolocation

During the initial data investigation, it became apparent that multiple latitude and longitude coordinates were associated with each customer zip code. Consequently, we opted to utilize the last recorded occurrence of each unique zip code along with its respective latitude and longitude. This decision was based on the premise that it represented the final location tracking of each parcel.

3.2.2. Merging of Datasets

1. Dataset for Demand Forecasting

For our demand forecasting, we've constructed a comprehensive dataset that integrates various aspects of our order data. This dataset (orders_forecasting) is created by merging information about the items ordered, customer reviews, and product details with our main orders dataset. This merge allows us to compile a multi-dimensional view of our orders, factoring in not only what was purchased and when but also customer feedback and product specifics.

2. Dataset for Route Optimisation

The dataset for route optimization (route_optimization) builds upon the demand forecasting dataset by incorporating geolocation data. By merging customer geolocation information, we enhance our demand forecasting dataset with spatial dimensions, enabling us to optimize delivery routes more effectively.

3.2.3. Handling Missing Values

We've encountered missing values along various columns. Upon assessing the missing values, we see that columns such as 'order_approved_at', 'order_delivered_carrier_date', and 'order_delivered_customer_date' have a substantial number of missing entries. This could indicate issues with order processing or data recording that merit further investigation. For features like 'product_photos_qty' and 'product_category_name', missing values could impact our understanding of product visibility and categorization, both of which are critical for demand forecasting.

However, not all features contribute equally to model performance. The decision to retain missing values and address them on a case-by-case basis allows us to tailor our data imputation strategies according to the importance of each feature within our models. For example, missing 'order_delivered_customer_date' values may be interpolated based on related timestamps, or we might use a model that can inherently handle missing values, like XGBoost. Meanwhile, for less impactful features, we could employ mean imputation, mode imputation, or even omit these records if their absence won't skew the overall data distribution.

4. Demand Forecasting

In the intricate landscape of demand forecasting, the deployment of machine learning models offers a promising avenue to glean actionable insights that can significantly enhance inventory management, marketing strategies, and customer satisfaction. Our exploration encompassed the utilization of Random Forest, Neural Network, and Extreme Gradient Boosting (XGBoost) models, each bringing its unique strengths and challenges to the task of predicting the total number of orders across various product categories.

4.1. Data Preparation

To start off the demand forecasting, we prepared the data by converting the datetime columns to pandas datetime type. This is to ensure that we can predict the orders on a monthly basis. Following which we checked for missing values and dropped them.

Data is aggregated to reflect the total number of orders per product category, and the resulting summary is merged back into the original dataframe to enrich it with a target variable for forecasting.

Features and the target variable are specified, and the dataset is split into training and test sets to facilitate model evaluation.

A preprocessing pipeline is established for both numerical and categorical features, ensuring the data is in a suitable format for the model.

4.2. Random Forest

The Random Forest model is highly effective for demand forecasting due to its ensemble approach that combines multiple decision trees to enhance predictive accuracy and prevent overfitting. Its ability to manage complex interactions between features makes it suitable for the multifaceted nature of demand

forecasting, which is influenced by various factors like pricing and seasonality. Moreover, Random Forest's inherent feature importance evaluation helps in identifying key drivers of demand, providing valuable insights for strategic inventory management. As a non-parametric method, it excels in modeling nonlinear relationships without predetermined assumptions, offering a versatile and interpretable tool for forecasting demand (Biswal, T, 2021).

Preprocessing Pipelines: Numeric and categorical features undergo distinct preprocessing steps to address missing values and scale or encode features accordingly, utilizing a ColumnTransformer. This tailored preprocessing ensures that the data is optimally prepared for modeling, handling different data types effectively.

Model Training and Prediction: The Random Forest Regressor, known for its versatility in regression tasks and ability to navigate complex data relationships with minimal hyperparameter adjustments, is trained on this preprocessed data. The model's performance is quantitatively evaluated using metrics like R-squared, RMSE, and MAE, which illuminate the model's fit and predictive accuracy on test data.

Feature Importances and Evaluation: Insight into the model's decision-making is gained through extracting feature importances, highlighting the most influential factors in predictions. A comparison between actual and predicted orders is visualized in a DataFrame, offering a direct view of the model's effectiveness. These evaluation steps provide a comprehensive assessment of the model's predictive power and accuracy, essential for refining and enhancing model performance.

4.3. Neural Network

Neural Networks are advanced tools for demand forecasting, well-suited for deciphering complex patterns within large datasets. Their deep learning capabilities are particularly effective in understanding the nuanced behaviors of consumer markets. In demand forecasting scenarios, Neural Networks shine where traditional models fall short. They adeptly handle a wide array of inputs, encompassing everything from past sales to evolving market trends, delivering a holistic view of demand influencers. Their scalability ensures that as more data becomes available, their precision improves, making them ideal for e-commerce platforms awash with copious amounts of data.

Training and Prediction: A neural network model is constructed using a sequential architecture with two hidden layers, leveraging the ReLU activation function for non-linearity and a single neuron output for regression. It's compiled with the Adam optimizer and mean squared error loss, trained over 100 epochs with batch size 32 and a 20% validation split to monitor performance. Post-training, the model predicts on test data, transforming these predictions to a comparable scale with the actual orders.

Feature Importance via Permutation: To discern the impact of each feature on the model's predictions, a permutation importance technique is applied. This involves systematically shuffling each feature in the test dataset and observing the effect on model accuracy (measured by mean squared error). The increase in error from the baseline (with no shuffling) indicates the feature's importance, with a higher increase signifying greater importance. This method provides a model-agnostic approach to evaluate feature relevance, offering valuable insights for model interpretation and further refinement.

4.4. XGBoost

XGBoost is recognised for its superb efficiency in demand forecasting, leveraging structured numerical data and the gradient boosting framework to provide accurate predictions. It's particularly adept at managing the complex and noisy datasets typical in consumer demand scenarios (practice, 2017).

Model Preparation and Hyperparameter Tuning: Our data was converted into the DMatrix format to facilitate fast and efficient model training. Initial model hyperparameters were thoughtfully selected, with a maximum tree depth of 6 and a learning rate of 0.3, to ensure a balance between learning complexity and the risk of overfitting. A systematic hyperparameter tuning was performed to fine-tune the model, focusing on improving the RMSE during cross-validation. This optimization process ensured that the final model was tailored to the unique characteristics of our dataset.

Model Training and Evaluation: After 100 rounds of boosting, aimed at sequentially improving upon the mistakes of previous iterations, the model was evaluated. Rather than focusing on traditional metrics alone, we placed a greater emphasis on the R-squared value for its ability to convey how well the model's predictions match actual demand variability.

4.5. Model Comparison for Demand Forecasting

In the critical task of demand forecasting, the choice of predictive modeling can have substantial implications for inventory management, customer satisfaction, and overall profitability. We have evaluated the three machine learning models using three key metrics: R-squared, Root Mean Square Error (RMSE) and Mean Absolute Error (MAE).

	R-squared	RMSE	MAE
Random Forest	1.0000	0.1142	0.0014
Neural Network	0.9996	21.9091	1.6230
XGBoost	0.9998	46.6835	30.4993

1. **Random Forest:**

R-squared: 1.0000 - This suggests that the model explains all the variability of the response data around its mean. In practice, a perfect R-squared is highly unusual and could indicate overfitting.

RMSE: 0.1142 - The RMSE is low, which suggests that the model has a good fit and the predictions are close to the actual values.

MAE: 0.0014 - A very low MAE indicates that the model performs well in terms of average magnitude of the errors in the predictions.

1. **Neural Network:**

R-squared: 0.9996 - This is a very high R-squared value, indicating that the model also captures almost all the variance in the data.

RMSE: 21.9091 - Compared to the Random Forest, the RMSE is significantly higher, suggesting larger errors between predicted and actual values.

MAE: 1.6230 - The MAE is higher as well, indicating that the average magnitude of the errors is greater than that of the Random Forest model.

2. **XGBoost:**

R-squared: 0.9998 - This is also a very high value, slightly better than the Neural Network, indicating excellent explanatory power.

RMSE: 46.6835 - This RMSE is quite high, suggesting that there are large errors between the predicted and actual values.

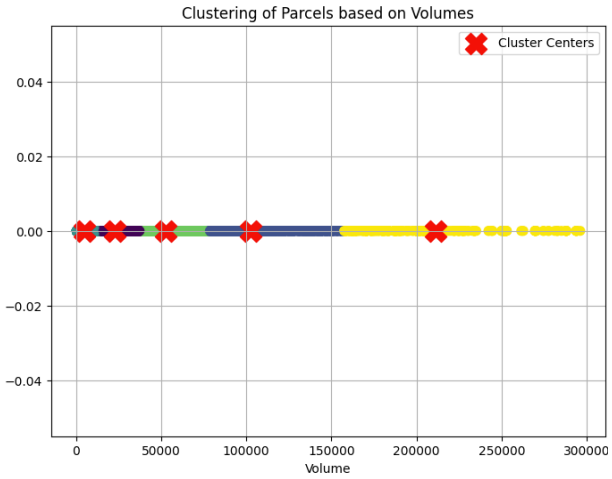
MAE: 30.4993 - This is by far the largest MAE, indicating the average magnitude of errors is quite high.

In evaluating the models for demand forecasting, the Random Forest model has shown a distinct advantage. Its R-squared value of 1.0000 indicates a theoretically perfect fit, which typically would be a concern for overfitting. However, the model's performance is also supported by the lowest RMSE and MAE among the candidates, suggesting it has not just learned the training data but can predict new data with high accuracy.While these results are promising, it is essential to conduct further evaluations through cross-validation or out-of-sample testing. This additional step will help to confirm the model's ability to generalize and its suitability for operational forecasting. Conclusively, the Random Forest model, based on the data provided, stands out as the most appropriate choice for demand forecasting due to its precision and reliability.

5. Route Optimisation

5.1. Clustering (volume)

The model was applied to the dataset containing product information. It aims to optimize the loading of a van by categorizing parcels based on their volumes and then efficiently allocating them to utilize the van's capacity.




```
Cluster Centers:
Cluster 1: 22808.186169778906
Cluster 2: 102596.84998086403
Cluster 3: 4713.353353777469
Cluster 4: 52855.723723015784
Cluster 5: 211177.80929095377
```

Figure 7: Clustering parcels based on volumes

Using K-means clustering, parcels are grouped into clusters based on their volumes, and the cluster centers are visualized. These cluster centers serve as reference points for better defining parcel categories and allow us to categorize different sizes of parcels into 5 categories: 'super small', 'small', 'medium', 'big', and 'super big'. Since parcels are assigned to categories based on their volumes, we would be able to determine the number of parcels in each category.

```
# Iterate through categories and allocate parcels proportionally
for category, count in parcel_counts.items():
    # Calculate the proportion of the van capacity for the current category
    category_proportion = count / sum(parcel_counts.values())
    # Calculate the volume of parcels to be loaded for this category
    category_volume = van_capacity * category_proportion
    # Update the loaded parcels with the rounded count
    loaded_parcels[category] = round(category_volume / categories[category])
    # Update the remaining capacity
    remaining_capacity -= loaded_parcels[category] * categories[category]
```

Figure 8: Finding the optimal combination of parcels

This code iterates through each category of parcels and allocates them proportionally based on the number in each category. It calculates the proportion of the van capacity that each category should occupy, then distributes the parcels accordingly, ensuring that the van's capacity is fully utilized. Hence, this optimization contributes to streamlined parcel management, ensures efficient parcel handling thus enhances delivery efficiency.

5.2. Clustering (location)

The clustering model was applied to the dataset containing geographical locations with the filtered data for São Paulo on 2017-05-05 due to its population density resembling that of Singapore's daily delivery demand.

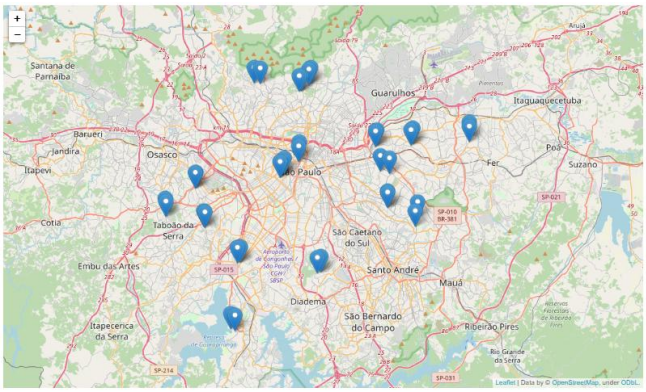


Figure 9: Data filtering to fit Singapore's context

The objective of the model was to identify spatial clusters, taking into account the specified number of clusters based on the constraint of available driver capacity for daily deployment determined by XGBoost. To achieve this, we employed the KMeans algorithm to partition the locations into clusters for each configuration.

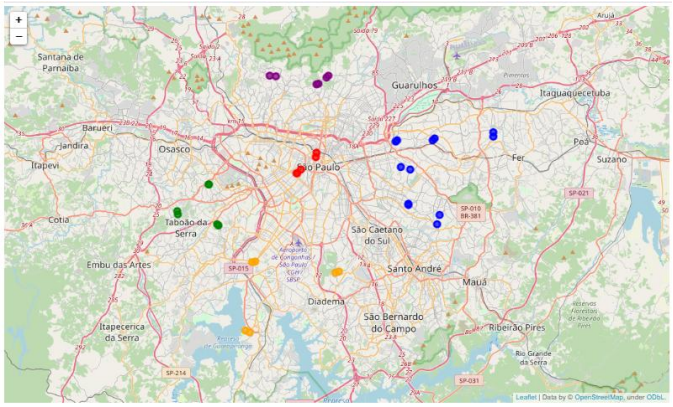


Figure 10: Clustering based on locations

The evaluation focused on calculating the average distance of each cluster, representing the compactness of the clusters. The average distance metric was computed as the mean distance from each point within a cluster to its cluster center. This evaluation method provides insights into the effectiveness of the clustering model in grouping spatially close points together.

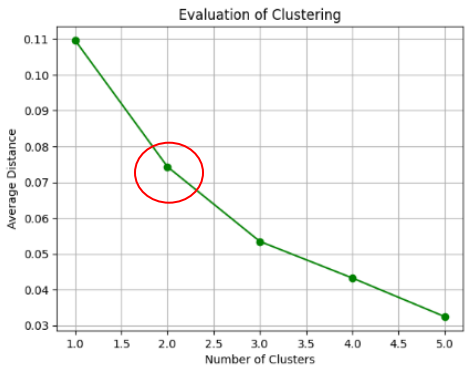


Figure 11: Evaluation of clustering

Firstly, the average distance between locations within each cluster was calculated. This metric aids managers in ensuring that the distance between various locations within each cluster is optimized for the dispatched drivers.

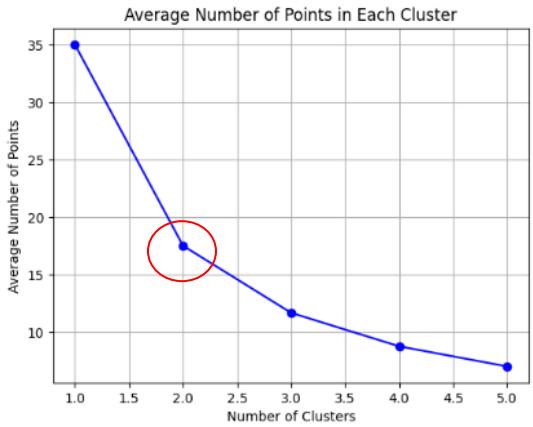


Figure 12: Average number of points in each cluster

Secondly, we determined the average number of parcels assigned to each cluster. This metric ensures that each driver has a manageable workload and that parcels can fit into the van for a single trip.

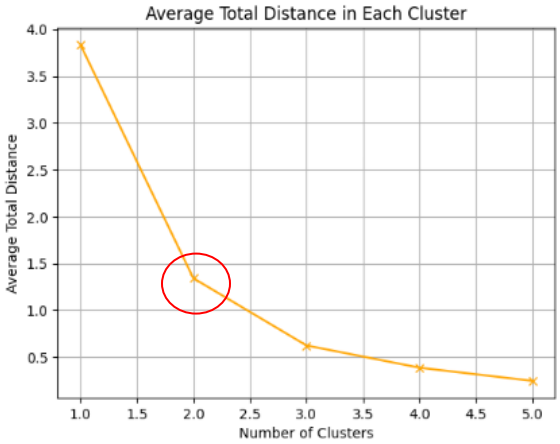


Figure 13: Average total distance in each cluster

Lastly, we computed the average total distance per cluster. This assists managers in ensuring that the overall distance covered by each driver is reasonable, allowing them to complete deliveries within a single shift.

The evaluation of the clustering model revealed valuable insights into the spatial distribution of the geographical locations. The model demonstrated varying levels of effectiveness in partitioning the locations into clusters based on different numbers of clusters. By analyzing the 3 evaluation metrics, we identified the optimal number of clusters that yielded the most compact clusters, is 2 clusters.

To fulfill the delivery order for this day, J&T will only need to send out 2 drivers to efficiently deliver all the parcels. This will help J&T with manpower allocation and scheduling. This information can be leveraged to enhance decision-making processes in various applications such as urban planning,

transportation optimization, and targeted marketing strategies. Overall, the clustering model serves as a valuable tool for spatial analysis and provides actionable insights for addressing real-world challenges related to geographical data.

5.3. Model Comparison for Route Optimization

By leveraging clustering to assign delivery destinations for each driver, we determine the optimal number of clusters that result in the most compact groupings. This approach facilitates efficient manpower allocation and scheduling for J&T. Moreover, it represents a promising strategy for optimizing J&T's delivery operations, ensuring timely parcel deliveries to customers while minimizing delays.

Clustering parcels works around the constraints that there is limited capacity for each van and this ensures efficient allocation of drivers each day, enabling all parcels to depart the hub immediately without requiring additional trips for those not accommodated in the initial delivery round. However, relying solely on parcel clustering may not entirely address route optimization and timely delivery challenges.

Therefore, the team chooses to utilize clustering to assign delivery destinations to solve delayed delivery. To enhance operational efficiency further, J&T should integrate parcel clustering into its decision-making processes. This integration ensures that the allocated number of drivers can accommodate all parcels in a single trip, minimizing the necessity for multiple returns to J&T's hub. By strategically incorporating clustering, J&T can streamline delivery processes and improve overall operational efficiency.

6. Proposed Solution

6.1. J&T Analytics

6.1.1. How It Works

J&T Analytics is an internal software platform where the management team can access real-time analysis on order demand forecasting and route optimisation planning. The software features include a monthly projected order quantity, monthly predicted number of drivers needed, daily parcel clustering and a manual or randomized driver allocation based on the clustering classification. The software predictions are continually updating and changing based on real-time data.

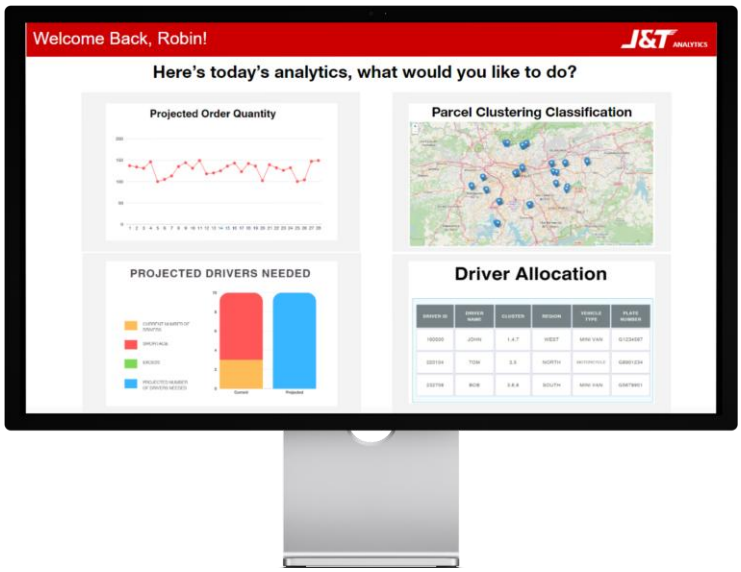


Figure 14: Home page of J&T Analytics

Figure 15: User interface of J&T Analytics App

Relevant information such as the clusters as well as driver allocation are reflected on the J&T Analytics app which is a mobile application exclusive to J&T delivery drivers. Drivers will be notified of the clusters they have been allocated to and the deliveries they have to make in the day. Using GPS technology, the app will recommend the fastest route by calculating the distance and time taken from the delivery hub to the respective clusters.

6.1.2. Impact on business

Empowered Employees and Enhanced Work Culture: Employees will benefit from an enhanced work culture where predictive scheduling and demand forecasting minimize uncertainty and overwork. This results in clearer expectations and a better balance between work and personal time, leading to improved job satisfaction. Furthermore, a culture that values proactive planning, backed by data, nurtures an environment of empowerment, where employees are more engaged and motivated to contribute to the company's success. As employees become ambassadors of innovation, this cultural shift can foster a strong

sense of loyalty and reduce turnover rates, reinforcing the company's foundation with a stable, skilled workforce.

Elevated Customer Experience: This high-touch customer service approach, underpinned by reliable delivery performance, will enhance the overall satisfaction and foster a sense of trust and reliability in J&T Express Singapore. By transforming the end-to-end delivery experience, the company is set to elevate its brand perception and create a competitive advantage that will resonate in the marketplace and generate increased customer loyalty and retention.

Operational Excellence: By automating key processes and employing route optimization, the company can expect to see a reduction in delivery times, increased accuracy in logistics planning, and a streamlined approach to parcel management. These improvements not only contribute to reducing operational overheads but also increase the company's capacity to handle a larger volume of orders. Additionally, with a more agile operational model, J&T can rapidly adapt to fluctuations in demand and scale operations up or down as needed without compromising service quality or employee well-being.

7. Business Valuation

The proposed solution's value is multidimensional, encompassing cost efficiencies, revenue enhancements, and qualitative benefits. The valuation encompasses not only the direct financial gains but also the strategic advantages that position J&T Express Singapore for future growth.

Cost savings and Increased Revenue: By refining driver schedules, we're set to reduce labor expenses by cutting down on overtime and the dependency on part-time staff during peak demand. Additionally, our route optimization strategy is projected to lower fuel and maintenance costs, thanks to reduced travel distances and minimized vehicle wear. Our solution directly contributes to revenue growth through increased operational throughput, enhancing our order handling capacity and facilitating higher fulfillment rates. Enhanced service quality is likely to foster customer retention, spawning repeat business and favorable word-of-mouth—key drivers of organic revenue growth. By delivering parcels on time consistently, we nurture customer loyalty and satisfaction, which are invaluable assets in the competitive e-commerce logistics market.

Market Growth and Expansion: The scalability of our solution ensures that J&T Express can manage increased order volumes effectively without a corresponding rise in resource allocation, enabling strategic expansion. Moreover, by automating and streamlining core processes, we anticipate a reduction in the probability of human error, thereby mitigating operational risks. The collective impact of these enhancements is a robust positioning of J&T Express Singapore as an industry leader, equipped with a competitive edge that not only meets but sets market standards. Data-driven insights will allow J&T to identify trends and capitalize on untapped market opportunities and emerging customer trends.

8. Feasibility

As J&T seeks to enhance its operational efficiency and customer satisfaction, the feasibility of our solution is important. This section evaluates the solutions viability across technical, financial, and operational dimensions, alongside its strategic fit with the industry's competitive and regulatory frameworks.

Seamless Technical Integration: The proposed solution is technically sound, leveraging proven data analytics and automation technologies for demand forecasting and route optimization. Its integration with existing operational frameworks is straightforward, requiring minimal disruption while offering seamless scalability to accommodate future growth. This ensures not just technical feasibility but also operational harmony with J&T Express Singapore's current and future needs.

Financial Justification and ROI: Initial investment requirements are offset by significant operational cost savings and revenue growth potential. The solution's financial model, highlighting reduced labor and operational costs alongside increased efficiency and customer retention, predicts a favorable return on investment (ROI).

Managed Risks: A phased rollout and comprehensive risk management plan, including contingency operations, ensure minimal risk during implementation. The solution's long-term impact extends beyond operational enhancements to solidify J&T Express Singapore's market position, promising sustained growth and increased brand equity.

9. Limitations

Resource Constraints and Scalability Challenges: Implementing the proposed solution entails significant upfront investment and poses scalability challenges as J&T Express Singapore grows. Balancing financial outlay with the expected benefits requires meticulous planning and justification. Moreover, as the company expands, ensuring the solution scales effectively to meet increasing demand without proportional increases in costs is crucial. Proactive financial management and designing the system for easy scalability will be vital.

Technical and Operational Integration Hurdles: The integration of new technologies presents technical challenges, from potential system incompatibilities to data security concerns. Additionally, there's the human element—ensuring staff adapt to and embrace new processes. A phased implementation approach, coupled with robust training and support, can mitigate these risks. Prioritizing data privacy and security from the outset will also be essential to protect sensitive information.

Market Dynamics and Customer Perception: The competitive landscape of the logistics industry means that staying ahead requires constant innovation and adaptation. Furthermore, there's the risk that customers may not immediately recognize or appreciate the improvements. To address this, continuous market analysis and flexible solution design will enable J&T Express Singapore to adapt to emerging trends and competitor strategies. Engaging customers through transparent communication and feedback mechanisms will help in refining the solution and bolstering customer satisfaction.

10. Recommendations

Foster Innovation and Adaptability

Commiting to a culture of continuous improvement and innovation. This encompasses regular updates to the solution with the latest technological advancements, expanding data analytics for sharper insights, and ensuring the system's architecture is scalable and flexible. Embrace emerging technologies like IoT and blockchain for enhanced operational efficiency and customer trust. Preparing for global market variations early will facilitate smoother expansions later.

Enhance Training and Customer Engagement

Deepen employee engagement with extensive training programs tailored to new technologies and operational changes, ensuring staff are well-prepared to maximize the benefits of the solution. Parallelly, develop robust mechanisms for customer feedback to continuously refine service offerings. Future interfaces should be more interactive and customer-centric, providing personalized experiences and transparent communication.

Strategic Partnerships

Exploring strategic partnerships can broaden the solution's capabilities and open up new avenues for innovation. This approach not only addresses environmental concerns but also strengthens the brand's appeal and market position. Moreover, Partnerships can help to alleviate the resource constraint as they would be able to get more funds, manpower and technology.

11. Conclusion

In conclusion, this report addressed J&T's challenges with delayed deliveries due to inefficient route planning and demand forecasting. We proposed a solution integrating demand forecasting models and route optimisation techniques to enhance operational efficiency and customer satisfaction. By leveraging data analytics, we not only tackled the immediate issues but also set the stage for more proactive and strategic decision-making.

APPENDIX

Appendix 1: Most Bought Items per month

```
Most bought items per month in 2017:
Month
1      furniture_decor
2      furniture_decor
3      furniture_decor
4       bed_bath_table
5       bed_bath_table
6       bed_bath_table
7       bed_bath_table
8       bed_bath_table
9       bed_bath_table
10      bed_bath_table
11      bed_bath_table
12      bed_bath_table
```

Top 1 item bought per month in 2017

Generally, customers would purchase items categorized under home furniture and housewares throughout 2017.

```
Most bought items per month in 2018:
Month
1                              bed_bath_table
2                      computers_accessories
3                              bed_bath_table
4                              bed_bath_table
5                               health_beauty
6                               health_beauty
7                               health_beauty
8                               health_beauty
9    kitchen_dining_laundry_garden_furniture
```

Top 1 item bought per month in 2017

The items bought in 2018 are different and have more variety compared to that of 2017. In the first quarter of the year, the patterns were similar to that in 2017 where the majority of products bought were from the houseware category, with the exception in the month of February where most customers bought computer accessories. In the second quarter of the year, health and beauty related items remained the most bought items.

Appendix 2: Data Cleaning

2.1 Geolocation

```
geolocation = pd.read_csv('/Users/Megan/Desktop/Y3S2/BC2407 Course Materials/project/data/olist_geolocation_dataset.csv')
geolocation = geolocation.groupby(['geolocation_city','geolocation_state','geolocation_zip_code_prefix']).last()
new_data = pd.read_csv('/Users/Megan/Desktop/Y3S2/BC2407 Course Materials/project/data/orders_forecasting.csv')
```

2.2 Merging of Datasets

```
# Merge datasets related to orders
orders_forecasting = orders \
    .merge(order_items_new, on='order_id', how='left') \
    .merge(order_reviews_new, on='order_id', how='left') \
    .merge(products_merged, on='product_id', how='left')
```

n

```
customers_geo = customers.merge(geolocation,
            left_on='customer_zip_code_prefix',
            right_on='geolocation_zip_code_prefix', how='left')
```

```
route_optimization = orders_forecasting.merge(customers_geo,
                                on='customer_id', how='left')
```

2.3 Handling Missing Values

```
# Find the number of missing values in each column
missing_values = orders_forecasting.isnull().sum()
print(missing_values)
```

```
order_id                            0
customer_id                         0
order_status                        0
order_purchase_timestamp            0
order_approved_at                 162
order_delivered_carrier_date     1980
order_delivered_customer_date    3253
order_estimated_delivery_date       0
order_item_id                     778
product_id                        778
price                             778
review_id                         961
review_score                      961
product_photos_qty               2390
product_category_name            2414
dtype: int64
```

Appendix 3: Demand Forecasting

3.1 Data Preparation

```
# Convert datetime columns to pandas datetime type
datetime_features = ['order_purchase_timestamp', 'order_approved_at',
                     'order_delivered_carrier_date', 'order_delivered_customer_date',
                     'order_estimated_delivery_date']
for feature in datetime_features:
    df[feature] = pd.to_datetime(df[feature])

# Drop rows with NaN values in columns that are crucial for analysis
df = df.dropna(subset=['product_category_name', 'price'])
```

```
# Drop rows with NaN values in columns that are crucial for analysis
df = df.dropna(subset=['product_category_name', 'price'])

# Aggregate to calculate total orders per product_category_name
df_agg = df.groupby('product_category_name').size().reset_index(name='total_orders')
```

```
# Prepare features (X) and target (y)
X = df.drop(['total_orders', 'order_item_id'], axis=1)
y = df['total_orders']

# Split the data into training and test sets
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2, random_state=123)
```

```
preprocessor = ColumnTransformer(transformers=[
    ('num', numeric_transformer, numeric_features),
    ('cat', categorical_transformer, categorical_features)
])

# Preprocess the training and test data
X_train_prepared = preprocessor.fit_transform(X_train)
X_test_prepared = preprocessor.transform(X_test)
```

3.2 Random Forest Results

```
Feature Importances:
                     Feature  Importance
17         x1_bed_bath_table    0.281792
53          x1_health_beauty    0.163871
49        x1_furniture_decor    0.117917
75         x1_sports_leisure    0.112690
25  x1_computers_accessories    0.107258
59             x1_housewares    0.074928
80          x1_watches_gifts    0.045222
79                   x1_toys    0.013856
15                  x1_auto    0.013136
78              x1_telephony    0.012996

Predicted vs Actual Orders:
   Actual Orders  Predicted Orders
0           4256            4256.0
1            240             240.0
2           9727            9727.0
3           4361            4361.0
4           8415            8415.0
5           8415            8415.0
6           4550            4550.0
7            605             605.0
8           6001            6001.0
9           3806            3806.0

Random Forest R-squared: 0.9999999988748949
Random Forest RMSE: 0.11420817938231197
Random Forest MAE: 0.0014487822349570213
```

3.3 Neural Network Results

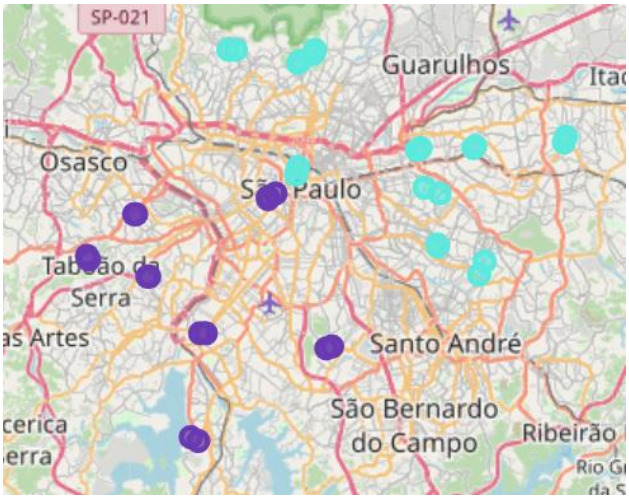
```
Top 10 Feature Importances from Neural Network:
product_category_name_bed_bath_table: 13565905.647194332
product_category_name_health_beauty: 8082234.170904282
product_category_name_sports_leisure: 5186711.702000941
product_category_name_furniture_decor: 4516311.781617027
product_category_name_computers_accessories: 3549704.073962774
product_category_name_housewares: 2201291.930010378
product_category_name_watches_gifts: 1103794.551734815
product_category_name_telephony: 269623.72035583726
product_category_name_garden_tools: 208287.90875101066
product_category_name_auto: 179064.10556451135

Neural Network Predicted vs Actual Orders:
   Actual Orders  Predicted Orders
0           4256       4255.437500
1            240        241.287292
2           9727       9727.199219
3           4361       4361.197266
4           8415       8414.630859
5           8415       8415.080078
6           4550       4550.375488
7            605        605.897156
8           6001       6001.078125
9           3806       3806.166748
Neural Network R-squared: 0.9999606719976316
Neural Network RMSE: 21.352641455902305
Neural Network MAE: 0.6353535660273024
```

3.4 XGBoost Results

```
XGBoost Feature Importance based on weight:
   Feature  Score
7      f17   13.0
15     f25   12.0
65     f75   12.0
39     f49   12.0
43     f53   12.0
49     f59   11.0
16     f26   10.0
36     f46   10.0
40     f50   10.0
24     f34   10.0

XGBoost Feature Importance based on gain:
   Feature         Score
7      f17  4.353940e+10
43     f53  2.775605e+10
39     f49  2.011848e+10
65     f75  1.927423e+10
15     f25  1.834663e+10
49     f59  1.372053e+10
70     f80  6.831567e+09
42     f52  1.482644e+09
68     f78  1.394791e+09
5      f15  8.593441e+08

XGBoost R-squared: 0.9998120137277017
XGBoost RMSE: 46.68353472598932
XGBoost MAE: 30.49934764577871

XGBoost Predicted vs Actual Orders:
   Actual Orders  Predicted Orders
0           4256       4229.745605
1            240        356.265747
2           9727       9709.041992
3           4361       4333.789062
4           8415       8404.477539
5           8415       8404.477539
6           4550       4523.685547
7            605        667.734070
8           6001       5978.281250
9           3806       3778.223389
```

Appendix 4: Route Optimisation

4.1 Clustering by location Results (Optimal)



REFERENCES

Singapore Business Review. (2024, March 31). J&T Express Singapore wins at SBR Technology Excellence Award for transportation management system. Retrieved from https://sbr.com.sg/co-written-partner/event-news/jt-express-singapore-wins-sbr-technology-excellence-award-transportation-management-system

Martech Asia. (2023, September 14). How J&T Express is transforming logistics and navigating challenges in the new retail landscape. Martech Asia.. https://martechasia.net/features/how-jt-express-is-transforming-logistics-and-navigating-challenges-in-the-new-retail-landscape/

Statista. (n.d.). Largest cities in Brazil. Retrieved March 31, 2024, from https://www.statista.com/statistics/259227/largest-cities-in-brazil/

practice. (2017, April 13). Beginners Tutorial on XGBoost and Parameter Tuning in R Tutorials & Notes Machine Learning. HackerEarth. Retrieved March 30, 2022, from https://www.hackerearth.com/practice/machine-learning/machine-learning-algorithms/beginners-tutorial-on-xgboost-parameter-tuning-r/tutorial/

Biswal, T. (2021, June 3). Random Forest for Time Series Forecasting. Analytics Vidhya. Retrieved from https://www.analyticsvidhya.com/blog/2021/06/random-forest-for-time-series-forecasting/