Optimizing Logistics Routes through Collaborative Recommendation Systems and Clustering Techniques: A Technological Approach to Aligning Planned and Actual Transportation Pathways

Hira Afzal
University of Trento
Trento, Italy
hira.afzal@studenti.unitn.it

Adnan Irshad
University of Trento
Trento, Italy
adnan.irshad@studenti.unitn.it

Afifa Saeed
University of Trento
Trento, Italy
afifa.saeed@studenti.unitn.it

Abstract—**In the domain of logistics management, the optimization of merchandise transportation routes across a network of cities presents a critical challenge, particularly in aligning the planned pathways with the actual routes taken by drivers. This project is designed to enhance the efficiency of goods movement using a company's fleet of trucks, by addressing the discrepancies observed between the intended and realized transportation routes. The core of our project methodology is the implementation of a Collaborative Recommendation System and clustering techniques, employing both item-based and user-based strategies to refine and personalize the standard transportation routes. Our approach leverages a utility factorization to map the interactions between drivers and transportation routes, utilizing K-means and K-medoid clustering algorithms to segment drivers based on historical data. This segmentation facilitates the generation of bespoke route recommendations that incorporate each driver's unique preferences and prior experiences, thereby aiming to minimize deviations from the pre-planned routes. Through a detailed comparative analysis employing cosine, Euclidean, and Jaccard similarity metrics, our solution offers tailored route suggestions that significantly reduce the gap between planned and actual route execution. The solution uses historical route data to construct detailed driver profiles, preferences and experience, enabling the generation of customized standard route recommendations to the company that enhance logistical operations and driver satisfaction. By incorporating sophisticated clustering algorithms and similarity metrics, the proposed solution assigns the best path to drivers based on their profiles.**

Index Terms—**Clustering Algorithms, Utility Matrix, Cosine, Euclidean, and Jaccard Similarities, Route Discrepancy Analysis, Data Mining, Synthetic Dataset.**

I. Introduction

The logistics company is confronted with the intricate challenge of streamlining the transportation routes for merchandise conveyance across diverse cities using its fleet of trucks. Each transit entails the movement of assorted merchandise types in specific quantities between city pairs, thereby constituting a comprehensive route. Presently, the company relies on suggested standard routes derived from specialized software, tailored to prevalent orders. However, a recurring issue emerges as drivers frequently peer off these prescribed routes due to personal inclinations or unanticipated circumstances, resulting in a notable variance between the planned and actual routes executed. In response to this discrepancy, the company seeks a technologically driven solution to address three pivotal objectives: offering strategic recommendations to the company regarding optimal standard routes grounded in historical route data analysis, facilitating the formulation of a meticulously ordered roster of standard routes for individual drivers, thereby mitigating route deviations, and crafting bespoke standard routes tailored to each driver's proclivities, to minimize discrepancies between planned and executed routes. Given the extant corpus of standard routes and historical data delineating the actual routes undertaken by drivers, the proposed system endeavours to optimize route planning processes, augment operational efficacy, and account for the idiosyncratic preferences and behaviours exhibited by drivers within the logistical framework.

Our solution employs a Collaborative Recommendation system and clustering, which encompasses two fundamental methodologies: item-based and user-based recommendation approaches. In our framework, the items correspond to transportation routes, while the drivers represent users. We establish a utility matrix where drivers are delineated as rows and routes as columns. Utilizing this matrix, we perform clustering using two prominent algorithms: K-means and K-medoids. This clustering facilitates two primary tasks: (i) creation of user profiles utilizing historical data derived from NFC and GPS tracking, and (ii) clustering to suggest updated standard routes for drivers to the company based on driver's experiences. Subsequently, we conducted a comparative analysis between actual and standard routes employing diverse metrics including cosine, euclidean, and Jaccard similarity matrice to analyze the route's divergence. This comparison enables us to tailor personalized recommendations for each driver, minimizing route deviations. Finally, we assigned a perfect route recommendation for each driver based on the driver's profiles created in task 1 with the least possible divergence.

In response to the limitations imposed by the unavailability of authentic datasets, our study initiated the development of a sophisticated synthetic dataset to simulate a broad range of logistical scenarios. This synthetic dataset serves as a foundational element for evaluating our proposed analytical framework, which integrates advanced data analysis and driver behavior modeling techniques.

Our research commenced with an in-depth exploration of clustering algorithms, focusing primarily on DBSCAN for its density-based clustering capabilities and K-means for its efficiency in partitioning data into distinct groups. The selection process ultimately favored K-means and K-medoids, with the latter being particularly valued for its use of actual data points as centroids, providing a tangible basis for cluster formation. The effectiveness of these algorithms was assessed using both Euclidean and cosine distance metrics to ensure comprehensive analysis. The evaluation phase involved comparing driver profiles against predefined standard routes using the selected distance metrics. This comparison highlighted a significant improvement in the correlation between drivers' profiles and their assigned routes, illustrating the robustness of our analytical approach. The final stage of our study focused on optimizing route assignments to align more closely with individual driver preferences. By maximizing the similarity between driver profiles and routes, our methodology demonstrates a nuanced approach to enhancing logistical efficiency through personalized route planning. This project contributes to the logistics and transportation field by presenting a method for route optimization that accounts for driver behaviors and preferences, as evidenced by our rigorous testing on the synthetic dataset.

II. Related Work

Logistics stands as one of the paramount activities in both personal and organizational realms, addressing the fundamental need for the movement of goods [1]. Goods encompass raw materials, semi-finished, or finished products, facilitating movement either between businesses within the same sector or across different sectors. Additionally, goods may traverse from manufacturers directly to end customers or via distributors or retailers. For decades, route optimization has remained a compelling subject of research. Often referred to as vehicle routing problems, this optimization endeavour aims to refine the itinerary of a fleet engaged in a round trip encompassing multiple stops [2]. Such trips commonly symbolize the distribution of goods within urban areas. Past investigations have demonstrated that route optimization holds the potential to substantially diminish transportation expenses, with potential cost savings ranging between 5% to 30% [3].

Mingbo et al. [4] introduce a systematic methodology for optimizing vehicle route planning, characterized by its efficiency and efficacy. Central to this novel approach is the utilization of multiple spanning trees as a criterion to classify customers into distinct sub-areas. Through an iterative process involving the selection of customers situated on the periphery of these sub-areas and their allocation to different spanning trees, more compact configurations with reduced distances are attained. Consequently, as the spanning trees serve as lower bounds for vehicle routes across all sub-areas, dispatching vehicles within each sub-area is facilitated. Rigorous simulation exercises have been conducted to empirically validate the efficacy of the proposed methodologies.

The Recommender Systems [5] discusses advanced matrix factorization techniques for enhancing recommendation systems, showcasing their superiority over traditional nearest-neighbour methods. It emphasizes incorporating additional data, such as implicit feedback and temporal effects, into the recommendation process. Integrating collaborative filtering alongside matrix factorization, we aim to refine our recommendation system, drawing parallels to the successes observed in consumer-based recommendation systems like Netflix.

Clustering plays a crucial role in analyzing and segmenting route data, allowing for the identification of common travel patterns and the optimization of the recommendation process. The project's application of clustering, particularly K-means, and k-medoids [6], [7], facilitates the grouping of similar routes, thereby enabling the generation of new standard routes that reflect real driving behaviors more accurately. This method enhances the system's ability to offer targeted and efficient route suggestions, directly addressing the project's objective to minimize discrepancies between planned and actual routes.

Anna Huang in her research [8] examines the efficacy of Jaccard and cosine similarity across various text document clustering scenarios, discovering that Jaccard similarity fares better with smaller datasets, whereas Cosine similarity excels in larger datasets. This approach, validated on a dataset encompassing over 6 million taxi trajectories [9], notably improves clustering results. K-medoids clustering, known for its robustness across different data types, selects k-medoids and iteratively reassigns data points to the closest medoid, optimizing intra-cluster distances until stability is achieved [10].

In summary, the proposed solution endeavors to address the intricate challenge of optimizing transportation routes within the logistics domain through the integration of collaborative recommendation systems, clustering techniques, and comparative route analysis. While commendable in its approach, the solution could benefit from a deeper exploration of advanced optimization algorithms, such as genetic algorithms [11], as demonstrated in existing literature. By drawing upon insights from diverse methodologies outlined in the literature, the proposed solution can evolve to meet the dynamic demands of modern logistics operations, ultimately striving toward greater efficacy and operational excellence.

III. Methodology

A. Dataset Generation

In the creation of datasets for our project, due to privacy constraints, it was necessary to fabricate data. This process entailed compiling a diverse array of information, including selecting from 46 Italian cities and 12 distinct types of goods, demanding extensive effort to ensure realism and variety in the synthetic data produced. The standard dataset was developed in alignment with hypothetical company directives, wherein routes were generated through random selection of 3 to 10 trips based on the geographic data. For each segment of these routes, specific details such as the origin, destination, and the goods being transported were fabricated, with each segment featuring a random selection of 2 to 4 types of goods in varying quantities. This procedure was repeated multiple times, creating a wide range of synthetic standard routes, subsequently archived in a JSON file to facilitate future analyses and insights into distribution network dynamics.

To simulate potential real-life deviations, instead of generating routes by random selection, we consider the distance between the cities. For example, if a route departure city is Trento and the destination city is Turin, then instead of generating random trips, we consider assigning the nearest cities from our cities database. To implement this strategy, we fetched the position coordinates from Wikipedia and generated a matrix containing each city's distance from all other cities in our database to make our routes more realistic and facilitate the realistic computation of distances between trips.

Actual routes introduce deliberate deviations from these standard routes, simulating real-world driver behaviours such as omitting certain cities, adding unplanned stops, replacing certain cities in a trip or varying the merchandise carried. This approach allows for a comprehensive simulation of logistics operations, capturing the complexity and unpredictability inherent in managing a fleet of trucks over a network of Italian cities. The dataset thus created, encompassing both standard and actual routes, reflects a nuanced spectrum of potential logistics scenarios, laying a robust foundation for the subsequent development and evaluation of a recommendation system aimed at optimizing route adherence and operational efficiency.

A dataset consisting of 30 standard routes and 760 actual routes with 1535 trips, consisting of 15 to 35 variations randomly against each assigned standard route for each 50 drivers has been generated to solve the project's problem statement.

B. Data Processor

The Route Data Processor provides a starting point for loading and preparing route datasets about the logistics sector, specifically focusing on the discrepancies between planned (standard) and actual routes taken by vehicles. It provides data structures to hold standard and actual route data, normalization techniques were applied to city names and merchandise quantities to ensure uniformity across the dataset. Additionally, the class was equipped with methods to generate sequences of cities and trips within routes, providing a structured representation of the data that reflects the actual logistics operations. It builds a comprehensive vocabulary and creates mappings for all unique entities in the dataset was driven by the requirement for data transformation. Textual data, such as city names and merchandise types, needed to be converted into numerical formats for processing by machine learning models. This approach facilitated the encoding and decoding of data, enabling seamless integration with subsequent analytical and model development phases. Key methods implemented in the class include:

- *Data Normalization:* Methods such as **normalize_city_names** and convert**_merchandise_quantities** were developed to standardize the data, addressing inconsistencies and formatting issues that could hinder analysis.
- *Sequence Generation:* The **get_route_cities_sequence** and get_trips_sequence methods were crucial for transforming raw route data into structured sequences that reflect the logistics operations more accurately.
- *Vocabulary Building:* The **_build_vocabulary** method systematically compiled a list of all unique entities (cities, merchandise, etc.) present in the dataset, laying the foundation for data encoding and decoding.
- *Mappings Creation:* The **_create_mappings** method established bidirectional mappings between textual and numerical representations of the vocabulary items, essential for data transformation tasks in machine learning workflows.

Overall, the Data Processor's structured approach forwards the pre-processing and analysis of logistics data. By providing a robust framework for data normalization, sequence generation, vocabulary building, and mappings creation, the class facilitates a deeper understanding of route planning and execution discrepancies.

C. Routes Encoding and Vectorization

The necessity to transform complex route data into a structured, analyzable format of matrices and vectors is critical for identifying patterns, discrepancies, and opportunities for optimization within the logistics routes data. The ultimate goal of this endeavour is to provide actionable insights that can lead to more efficient and reliable logistics operations. The RouteVectorizer class is a pivotal component in the progression of a project aimed at enhancing route representation for the machine models. This transformation is facilitated by the following methods:

- *_encode_data:* Encodes the route data by converting city names and merchandise types into their corresponding numerical IDs, leveraging the vocabulary built during the preprocessing stage. In developing the encoding strategies for the project, an exploration of various methods highlighted the limitations of binary encoding in capturing essential route details, such as the sequence of cities visited and the quantities of merchandise transported. This insight led to the adoption of One-Hot Encoding, a method that preserves the integrity of the route's structural and quantitative data. This encoding approach, by meticulously detailing each trip and then aggregating them to form a complete route, ensures the preservation of merchandise significance throughout the journey, offering a nuanced data representation suited for in-depth analysis.
- *_vectorize_data:* Transforms the encoded route data into a structured numerical format, either as matrices or vectors,

depending on the specified vectorization technique. This transformation is crucial for feeding the data into machine learning models.

- *_minhash_vectorize_data:* Generates Minhash signatures for the vectorized routes, facilitating efficient similarity search and clustering operations, which are essential for analyzing patterns in route data.

Overall route encoding and vectorization workflow integrates the functionalities of the RouteDataProcessor component to represent raw data in a structured format conducive to advanced analysis. Starting with the translation of raw data into numerical identifiers through the RouteDataProcessor class, the process advances with the Encoder transforming these encoded routes into matrices and vectors.

D. Collaborative Filtering Recommendation System

In the realm of route optimization and driver behaviour analysis, in this project, we leverage collaborative filtering techniques, specifically focusing on a hybrid approach that incorporates both user-based and item-based filtering, facilitated through matrix factorization methods. The overarching goal was to enhance route planning efficiency, minimize deviations, and cater to driver preferences, ultimately leading to improved operational efficiency and reduced logistics costs.

Content-based Recommendation recommends items similar to what a user has liked in the past. It relies heavily on the features of the items themselves. In this case, about each route, such as the types of cities or locations included, specific characteristics of each trip, road conditions, scenic value, and so on. However, in our data, we don't have any such features. Therefore, we move with the collaborative filtering recommendation system which is designed to harness the collective experiences and preferences of drivers regarding their route choices. Recognizing that drivers with similar route preferences in the past are likely to exhibit similar preferences in the future, the system aims to recommend routes that a driver might prefer based on the preferences of similar drivers.

- *Step 1: Constructing the Driver-Route Matrix*
  At the heart of the recommendation system is the Matrix Factorization class, which creates a utility matrix representing the interactions between drivers and routes. This matrix serves as the foundation for extracting insights into driver preferences and route characteristics. The matrix factorization technique was chosen for its efficacy in identifying latent factors that influence route preferences, thus enabling a nuanced understanding of driver-route dynamics
  This matrix assigns drivers to rows, each representing their engagement with various trips, while columns denote individual trips, characterized by their standard routes. The matrix is populated by contrasting the actual routes taken by drivers against these standard routes, assessing compliance and deviations. This assessment could consider factors like adherence to prescribed routes, efficiency, and personal route preferences, thereby quantifying the degree of alignment between driver actions and expected norms.
- *Step 2: Ranking Routes based on Driver Profiles*
  The second step aims to enhance the matrix and rank the routes by integrating driver profile similarities, thereby improving the system's capability to forecast preferences for routes not previously encountered by drivers. *TripsRanking* class designed to rank trips based on a driver-route matrix, leveraging the principles of item-based collaborative filtering. The objective is to distil actionable insights from the utility matrix—comprising trip ratings and driver interactions—to inform and optimize trip recommendations. By evaluating and ranking trips according to their relevance and popularity among drivers, the system aims to personalize route suggestions, thereby aligning route planning with driver preferences and historical patterns.

Utility matrices are the backbone of collaborative filtering within recommendation systems, crucial for delineating and prognosticating user-item interactions. This case specifically addresses driver-trip interactions, employing a collaborative filtering framework to predict driver behaviours towards unexplored routes. The inclusion of driver profile similarities significantly bolsters the utility matrix's predictive accuracy, facilitating a tailored and dynamic recommendation system adept at accommodating the diverse preferences and behaviours of drivers.

E. Clustering and Clusters Analysis

This module signifies a strategic advancement in leveraging clustering algorithms to enhance route optimization for logistics operations. *ClusteringAnalyzer* analyze and categorizes routes into clusters, finding the best cluster counts and hyperparameters tuning for Kmeans++ and K-medoids algorithms facilitating the identification of inherent patterns and similarities within the route data. The objective is to employ clustering techniques, such as KMeans or KMedoids, to discern groups of routes that exhibit similar characteristics.

F. Task-1: Standard Routes Recommendation to Company

One of the primary objectives of the project is to recommend and generate standard route recommendations for the company based on actual routes. This is done by applying clustering on driver-routes vector representation of routes representing a novel approach to enhancing logistics operations. The process involves analyzing vectorized route data using clustering algorithms (KMeans and KMedoids), determining cluster centroids, and constructing recommended routes based on these centroids. This approach aims to consolidate insights from existing route data to propose standardized routes that align with prevalent driving patterns and preferences.

The methodology for generating standard routes recommendations encompasses several key steps:

- *Clustering Analysis:* Utilizing the ClusteringAnalyzer class, routes are clustered using both KMeans and KMedoids algorithms. This clustering segregates the routes

Algorithm 1 Matrix Factorization for Utility Matrix Generation

```
function CREATE_UTILITY_METRICS(self)
    $sum\_matrix\_, count\_matrix\_ \leftarrow$ zero matrices of size$(self.num\_drivers, self.num\_trips)$
    $vec\_length \leftarrow self.num\_cities + self.num\_merchandise$
    for each row in $self.df\_$ do
        for $act\_index, act\_trip\_seq\_$ in enumerate(row['trips_sequence_act']) do
            $act\_trip\_id \leftarrow self.routes\_vocab\_.get\_trip\_id(act\_trip\_seq\_)$
            $cum\_similarity, comp\_count \leftarrow 0, 0$
            for $k$ from 0 to len(row['trips_sequence_std']) - 1 do
                $std\_vec\_slice \leftarrow row[self.vecs\_col +' \_std'][k * vec\_length : (k+1) * vec\_length]$
                $act\_vec\_slice \leftarrow row[self.vecs\_col +' \_act'][k * vec\_length : (k+1) * vec\_length]$
                $similarity\_ \leftarrow$ cos_sim$(act\_vec\_slice, std\_vec\_slice)[0, 0]$
                if $similarity\_ == 1$ and $act\_index == k$ then
                    $cum\_similarity, comp\_count \leftarrow 1, 1$
                else
                    $cum\_similarity, comp\_count \leftarrow cum\_similarity + similarity\_, comp\_count + 1$
                end if
            end for
            $driver\_id \leftarrow self.routes\_vocab\_.get\_driver\_id(row['driver'])$
            $sum\_matrix\_[driver\_id, act\_trip\_id] \leftarrow sum\_matrix\_[driver\_id, act\_trip\_id] + cum\_similarity$
            $count\_matrix\_[driver\_id, act\_trip\_id] \leftarrow count\_matrix\_[driver\_id, act\_trip\_id] + comp\_count$
        end for
    end for
    $mask, u\_matrix\_ \leftarrow count\_matrix\_ \neq 0,$ zero matrix of same size as$sum\_matrix\_$
    $u\_matrix\_[mask] \leftarrow$ round$(sum\_matrix\_[mask]/count\_matrix\_[mask], 5)$
    return $u\_matrix\_, count\_matrix\_, sum\_matrix\_$
end function
```

Algorithm 2 Predict Ratings for Drivers and Routes

```
function PREDICT_RATINGS(self, driver_route_metric, count_matrix_, top_k_drivers, precision_=5)
    $predicted\_ratings\_ \leftarrow$ copy of $driver\_route\_metric$
    for $driver\_1$ from 0 to $self.num\_drivers - 1$ do
        $top\_k\_similar\_drivers\_ \leftarrow self.\_get\_top\_k\_similar\_drivers(driver\_route\_metric, top\_k\_drivers)$
        for $trip\_id$ from 0 to $self.num\_trips - 1$ do
            if $count\_matrix\_[driver\_1, trip\_id] == 0$ then
                $rating, similarity \leftarrow 0, 0$
                for $driver\_2$ in $top\_k\_similar\_drivers\_[driver\_1]$ do
                    if $self.driver\_route\_metrics[driver\_2, trip\_id] > 0$ then
                        $rating \leftarrow rating + driver\_route\_metric[driver\_1, driver\_2] \times self.driver\_route\_metrics[driver\_2, trip\_id]$
                        $similarity \leftarrow similarity + driver\_route\_metric[driver\_1, driver\_2]$
                    end if
                end for
                if $similarity > 0$ then
                    $predicted\_ratings\_[driver\_1, trip\_id] \leftarrow$ round$(rating/similarity, precision\_)$
                end if
            end if
        end for
    end for
    return $predicted\_ratings\_$
end function
```

into distinct groups based on similarity in route characteristics.

- *Centroid Analysis:* For each cluster, the centroid is calculated to represent the 'average' route within the cluster.

This centroid serves as a basis for constructing recommended standard routes, embodying the commonalities among routes within the cluster.

- *Route Reconstruction:* Based on the centroids, recommended routes are constructed by translating the centroid vector back into route and merchandise information. This process involves decoding the vectorized route data to identify starting and ending cities for each trip within a route, as well as the types and quantities of merchandise involved.

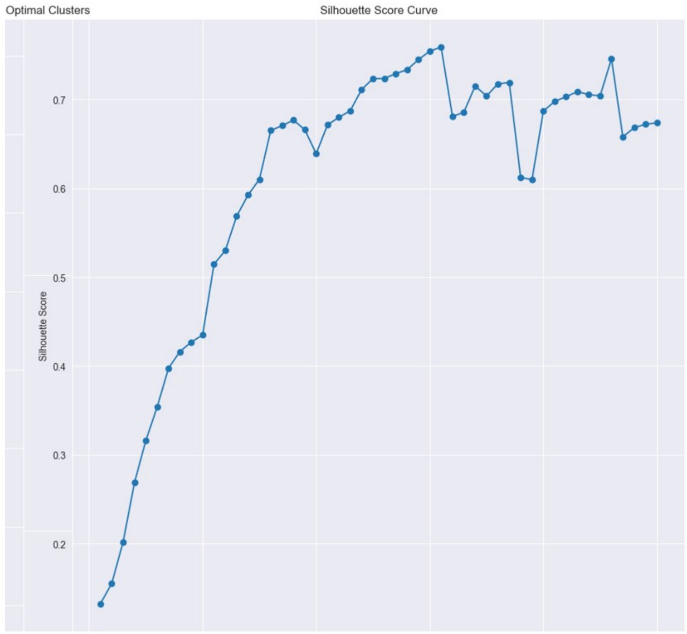

Fig. 1. Finding Clusters Count Based on Silhouette Curve

The analysis yielded a set of recommended standard routes, derived from the clustering of existing route data. These recommendations were categorized based on the clustering algorithm used (KMeans and KMedoids), with a combined total of recommended routes presented to the company for consideration.

G. Task-2: Driver Specific Less Divergence Routes

The objective of task 2 is to assign the list of standard routes to each driver with fewer divergence routes in a way that the topmost route is the one with the least divergence from the driver's actual route. The approach combines the newly recommended standard routes from task-1, combine with the old standard routes and measures the similarity measures between vectorized recommended standard routes and actual routes based on various metrics, including cosine similarity, euclidean distances, and Jaccard similarity. These scores provide a quantitative basis for assessing the degree of alignment between recommended and actual routes.

The similarity scores are then aggregated to compute an overall divergence statistic for each standard route relative to the drivers' actual routes. This analysis takes into account both the mean similarity scores and the frequency of route selections, normalizing these factors to derive a comprehensive measure of route preference.

Based on the computed divergence statistics, the top-k standard routes exhibiting the least divergence from each driver's actual routes are identified. This selection process prioritizes routes that not only align closely with historical route choices but also reflect a consensus among drivers regarding preferred routes.

H. Task-3: Driver's Ideal Route

The goal is to generate an ideal route for each driver which is the least divergence derived from item-based collaborative filtering, which assesses the drivers' historical preferences, ratings, or experiences with specific routes. This methodology encompasses several key steps:

- *Route Ranking Analysis:* Utilizing a matrix that ranks routes based on a driver-route matrix for item-based collaborative filtering, the method evaluates routes based on the drivers' historical preferences and experiences.
- *Optimal Route Construction:* For each driver, the system iteratively selects the best-rated trips to construct an optimal route. This selection process prioritizes trips that seamlessly connect, ensuring a coherent and efficient route from start to finish.
- *Route Recommendation:* The final step involves compiling the constructed optimal routes into a list of dictionaries, each containing the recommended optimal route for a specific driver. This recommendation considers both the route's overall rating and its logistical feasibility.

By focusing on the alignment between recommended routes and drivers' historical preferences, this approach promotes the adoption of routes that are both efficient and satisfying for drivers. The insights generated through this process contribute to the broader goal of optimizing route planning, reducing operational inefficiencies, and improving driver satisfaction through personalized route recommendations.

I. Evaluation

- Overview of Generated Routes and Similarity Metrics: The analysis of the project's outcomes reveals significant insights into the optimization of transportation routes within the logistics domain. The project successfully generated a total of 79 recommended standard routes for the company, identified 50 top-k least divergence routes for drivers, and constructed 50 optimal routes tailored to individual driver preferences. These outcomes highlight the effectiveness of the collaborative recommendation system and clustering techniques in personalizing route recommendations.
- Similarity Analysis between Old Standard and Actual Routes: The initial similarity analysis between old standard and actual routes utilized Jaccard, cosine, and Euclidean metrics to evaluate the alignment between planned and executed routes. The Jaccard similarity

matrix, calculated for a subset of the data, yielded a mean similarity score of 0.5306 with a standard deviation of 0.1216, indicating moderate alignment. However, the cosine similarity matrix across a larger dataset presented a lower mean similarity of 0.0521, suggesting a greater divergence in route execution relative to the planned routes. The Euclidean similarity matrix further corroborated this divergence with a high mean value of 42.8608, emphasizing the discrepancies between planned and actual routes.
- Similarity Analysis between New Standard and Actual Routes: The subsequent analysis, focusing on the new standard versus actual routes, demonstrated an adjustment in similarity metrics. The Jaccard similarity for the new standard against actual routes showed an improved mean score of 0.5 with a lower standard deviation of 0.0195, reflecting a more consistent alignment across routes. The cosine similarity mean score slightly decreased to 0.0466, with a reduction in standard deviation to 0.1117, indicating a more uniform but still divergent set of route executions. The Euclidean similarity measure increased to a mean of 46.6016, suggesting that while the new standard routes might reflect a closer grouping of route characteristics, the quantitative divergence remains significant.

Conclusion

In this project, a system was designed to optimize and recommend transportation routes, focusing on minimizing the discrepancies between planned and actual routes undertaken by logistics drivers. Leveraging advanced data mining techniques, including collaborative filtering and clustering methods such as KMeans and KMedoids, the system endeavours to produce standard route recommendations that resonate more closely with the historical patterns and preferences of drivers. The utilization of clustering algorithms was instrumental in formulating new standard routes that better encapsulate the drivers' actual route preferences. Moreover, the application of collaborative filtering techniques within the utility matrices further honed these route recommendations, ensuring they were both relevant and personalized. The primary challenge encountered in this project was finding and generating datasets and evaluating the quality of the newly recommended routes. This evaluation was crucial for refining the recommendation process and accurately interpreting the outcomes. However, the methodologies' effectiveness in accurately assessing the quality of the route recommendations remains an area for further exploration, especially concerning the ideal standard routes formulated for individual drivers in the third task of the study.

In conclusion, the developed solutions showcase a dynamic and flexible approach to addressing route optimization challenges, with the capability to test various techniques and metrics, extending beyond the Jaccard and cosine similarity measures. The process of identifying optimal clustering configurations and generating new standard routes highlights the system's versatility, marking it as a foundational tool for route optimization and recommendation efforts. This flexibility and adaptability signify the system's potential to serve as a comprehensive solution in the ongoing pursuit of enhancing route planning and execution within the logistics sector.

References

[1] N. Viswanadham and R. Gaonkar. E-logistics: trends and opportunities. E-Logistics Research WP TLI-AP/01/01, The Logistics Institute Asia-Pacific, Jan 2001.

[2] D. Cattaruzza, N. Absi, D. Feillet, and J. Gonzalez-Feliu. Vehicle routing problems for city logistics. *EURO Journal on Transportation and Logistics*, 6(1):51–79, 2017.

[3] G. Hasle, K.-A. Lie, and E. Quak. *Geometric Modelling Numerical Simulation and Optimization*. Springer, 2007.

[4] M. Zhao, T. W. S. Chow, and K. F. Tsang. Vehicle route planning for logistics network optimization via multiple spanning tree. In *2016 IEEE 14th International Conference on Industrial Informatics (INDIN)*, pages 800–805, Poitiers, France, 2016.

[5] Yehuda Koren, Robert Bell, and Chris Volinsky. Matrix factorization techniques for recommender systems. *IEEE Computer Society*, August 2009. ISSN 0018-9162/09/$26.00 © 2009 IEEE.

[6] Ma Haonan, Yong Chen He, Meng Huang, Yana Wen, Yuhan Cheng, and Yifei Jin. Application of k-means clustering algorithms in optimizing logistics distribution routes. In *2019 6th International Conference on Systems and Informatics (ICSAI)*, pages 1466–1470. IEEE, 2019.

[7] Abiodun M Ikotun, Absalom E Ezugwu, Laith Abualigah, Belal Abuhaija, and Jia Heming. K-means clustering algorithms: A comprehensive review, variants analysis, and advances in the era of big data. *Information Sciences*, 622:178–210, 2023.

[8] H. . Anna. Similarity measures for text document clustering. In *Proceedings of the sixth New Zealand Computer Science Research Student Conference (NZCSRSC2008)*, volume 4, pages 9–56, 2008.

[9] Y. He, F. Zhang, Y. Li, J. Huang, L. Yin, and C. Xu. Multiple routes recommendation system on massive taxi trajectories. *Tsinghua Science and Technology*, 21(5):510–520, 2016.

[10] U. K. Kaur, Noor Kamal, and D. Singh. K-medoid clustering algorithm—a review. *Int. J. Comput. Appl. Technol*, 1(1):42–45, 2014.

[11] N. Uddin, H. Hermawan, N. L. Rachmawati, and H. Tannady. Genetic algorithm for logistics-route optimization in urban area. In *2022 IEEE World AI IoT Congress (AIIoT)*, pages 071–076, Seattle, WA, USA, 2022.