|  |
| --- |
| TECHNICAL REPORT |

|  |
| --- |
| Electrical & Computer Engineering & Computer Science (ECECS) |

|  |  |
| --- | --- |
| SEMESTER II |  |



Contents

[Project Name 2](#_Toc96341549)

[Executive Summary 2](#_Toc96341550)

[Technical Report 3](#_Toc96341551)

[Highlights of Project 3](file:///C:\Users\ardia\Desktop\TEACHING\SPRING%2022\DSCI6002%20Data%20Science\Projects\Technical%20report%20Template.docx#_Toc96341552)

[Submitted on: 3](file:///C:\Users\ardia\Desktop\TEACHING\SPRING%2022\DSCI6002%20Data%20Science\Projects\Technical%20report%20Template.docx#_Toc96341553)

[Abstract 4](#_Toc96341554)

[Methodology 5](#_Toc96341555)

[Results Section 5](#_Toc96341556)

[Discussion 6](#_Toc96341557)

[Conclusion 6](#_Toc96341558)

[Contributions/References 7](#_Toc96341559)

|  |
| --- |
| Project Name |

|  |  |  |
| --- | --- | --- |
| Executive Summary This study details the creation and implementation of a machine learning model for predicting rail ticket costs. The model forecasts ticket prices based on many input variables by utilizing a linear regression technique, which is enabled by Amazon SageMaker. | | |
| person at a table writing in a notebook with people around | | |
| **Team Members:** Bhavana Yelisetti Apurva Ravikar Chavan Kavitha MadirajuLalitha Sravani Gannavarapu | **Questions?**  Contact : |  |

|  |
| --- |
| Technical Report |

|  |  |
| --- | --- |
| **TRAIN TICKET PRICE PREDICTION** |  |
| Highlights of Project Predicting train ticket pricing is an attempt to bring together state-of-the-art technology with useful economic applications. This study attempts to provide important insights into pricing patterns in the transportation sector by maximizing model performance, utilizing cutting-edge machine learning technologies, and carefully preparing data.  The project's core value is its dedication to superior data preparation. To guarantee uniformity and consistency, raw training data that was taken from an S3 bucket was subjected to stringent cleaning and standardization procedures. Precisely designed custom parsing methods were used to remove non-numeric characters, providing a strong basis for further examination. The dataset was prepared for numerical exploration by converting all of the columns to float data type. This allowed for well-informed decision-making throughout the modeling process. Submitted on: |

SUBMITTED ON: April 22nd,2024

## Abstract

The core and importance of a project aimed at employing cutting-edge machine learning techniques to estimate rail ticket prices are captured in this abstract. The project starts with painstaking data preprocessing, using unique parsing methods and numerical conversion to guarantee the training dataset's quality and consistency. The project smoothly moves into model building by employing Amazon SageMaker and linear regression to anticipate ticket prices as a continuous variable. In order to optimize predicted accuracy and scalability while fine-tuning model performance, hyperparameter optimization is essential. The created model has the potential to significantly impact business in the transportation sector by allowing stakeholders to better allocate resources, optimize revenue management techniques, and raise customer happiness. In the future, the project will pave the way for additional developments in predictive analytics, encouraging creativity and data-driven.

Review of available research

Many studies that examined the application of machine learning techniques to forecast rail ticket prices focused mostly on regression analysis. With positive forecast accuracy findings, Wang et al. (2019) used demographic data and historical ticket sales data to construct a pricing model based on gradient boosting regression. Liu et al.'s (2020) use of ensemble learning techniques to anticipate ticket prices highlights the importance of feature engineering and model selection in enhancing predictive performance.

Researchers have also looked at the impact of external factors on the dynamics of ticket prices in addition to machine learning methods. Chen et al.'s (2018) analysis of the relationships between fuel prices, seasonality, and competition and their effects on train ticket pricing revealed complex relationships requiring advanced modeling methods. Furthermore, studies have looked at how dynamic pricing systems work to control demand and maximize income utilizing information from fields like economics and operations research (Clemons & Gu, 2017).

Although prior research provides valuable insights into the prediction of train ticket pricing, some gaps and unexplored research topics remain. First off, real-time data streams combined with advanced analytics techniques show promise for increasing the timeliness and accuracy of price estimates, especially under dynamic market settings. Pricing models can also be improved and made more flexible by including consumer preferences and behavior data. This enables the creation of unique price plans for individual visitors.

## Methodology

## 1. Data Collection:

Obtaining past train ticket sales data from dependable sources, including ticketing platforms or transportation organizations, is the initial stage in the process. A wide variety of factors, such as ticket costs, travel dates, routes, train kinds, and passenger demographics, should be included in this dataset.

II. Business Understanding:

The objective of the project is to create a predictive model for train ticket prices in order to help transportation businesses improve customer happiness and revenue management tactics. Stakeholders can make educated judgments about pricing, resource allocation, and operational planning by precisely forecasting ticket prices.

III. Data Preparation:

To guarantee data quality and consistency, the gathered dataset is carefully preprocessed. To reduce the possibility of biases in the analysis, duplicates are eliminated, missing values are handled, and outliers are dealt with. To make categorical variables compatible with machine learning algorithms, encoding methods like label encoding and one-hot encoding are used.

IV.Modeling:

Because of its ease of use and interpretability, linear regression is the main modeling technique used in this study. For contrast, ensemble techniques like random forest or gradient boosting could be investigated. To better identify patterns and correlations in the data, feature engineering techniques are used to generate new features or modify current ones.

V.Evaluation:

An independent test set is used to evaluate the model's performance in order to gauge its capacity for generalization. The predicted accuracy of the model is measured using common evaluation metrics including mean absolute error (MAE), mean squared error (MSE), and root mean square error (RMSE). Extra information about the model's performance can be obtained by visually comparing the projected and real ticket prices.

VI. Iterative Refinement:

The predictive model can be continuously improved and refined because to the iterative approach of the methodology. Subsequent versions of the modeling process are guided by insights gleaned from monitoring and feedback from stakeholders. Methods like cross-validation and hyperparameter tuning are used to make sure the model is robust and performs well.

## Results Section

1. Data Engineering Pipeline:

Data Ingestion:

The process of gathering and integrating data into a centralized repository from multiple sources is known as data ingestion. Python packages like NumPy and Pandas were used for this project's data extraction and manipulation. Large datasets kept in cloud environments were also managed and accessed with the use of solutions like AWS S3 and AWS Glue.

Data Storage:

Relational database management systems (RDBMS), such PostgreSQL or MySQL, were used to store the gathered data in an organized manner. Strong storing capacities and effective data querying for analysis and modeling are provided by these databases.

Data Processing:

The procedures involved in cleaning, converting, and getting the data ready for analysis are all included in data processing. For data preprocessing tasks including handling missing values, encoding categorical variables, and scaling numerical features, Python packages like Pandas and Scikit-learn were utilized.

Data Consumption:

To enable real-time predictions, the trained predictive model is implemented in a secure and scalable environment. Machine learning models can be hosted as web services or APIs with the help of tools like Amazon SageMaker. This makes it possible to incorporate the predictive model into current business procedures or apps with ease.

Model Deployment:

The cloud-based systems like Amazon SageMaker are used to generate the model's deployable environment. These systems provide scalability, dependability, and security characteristics in their managed model deployment services. End users may easily access and utilize the model because it is delivered as a web service and enclosed in a containerized environment.

Data Visualization:

To aid in comprehension and decision-making, the prediction model's output is presented through detailed visualizations. To create visualizations like scatter plots, line charts, and heatmaps, Python tools like Matplotlib, Seaborn, and Plotly are used. By highlighting correlations, trends, and patterns in the data, these visualizations improve comprehension of the model's predictions.

Deployment:

## After the model is put into use, it is constantly observed and maintained to guarantee peak efficiency and dependability. In order to identify and address any potential problems during operation, automatic alerts and notifications are also put up.

## Discussion

Our study set out to estimate train ticket prices using a thorough data exploration process, rigorous model construction, and perceptive result interpretation. The preceding sections' findings make it clear that, although our study offers insightful information, it also highlights the subtleties and complexity of price dynamics in the transportation industry.

Finding patterns and links in historical train ticket sales data was the main objective of our research topic, which aimed to create a prediction model that would help transportation businesses improve customer happiness and revenue management tactics. We carefully designed a data pipeline using the CRISP-DM approach as a guide. We used cutting-edge technologies and methods to ingest, store, process, and use data in a useful and useful way.

Descriptive statistics and visualizations based on our analysis's findings highlight important trends and patterns in the dataset. But it's important to recognize that, although being a useful tool for predicting ticket prices, our model has some drawbacks. Like any predictive model, it has assumptions and inherent uncertainties that should be carefully considered.

It is crucial to acknowledge that, despite our greatest efforts, our study might only provide a piece of the complicated puzzle that is the pricing dynamics in the transportation industry. There are several variables that affect ticket prices, such as competition pricing strategies, market demand, and external economic considerations. These aspects make modeling and prediction extremely difficult.

Though we have made every attempt, we must acknowledge that our study might only provide a piece of the complicated puzzle that is the pricing dynamics in the transportation industry. Modeling and prediction are extremely difficult due to the wide range of factors that affect ticket prices, such as competition pricing strategies, market demand, and external economic variables.

However, our research constitutes a noteworthy advancement in comprehending and tackling these difficulties. We have gained important insights that help guide operational planning and strategic decision-making in transportation organizations by utilizing machine learning techniques and data-driven approaches.

In order to further improve and hone our predictive model going future, we must take a comprehensive and iterative approach. This calls for ongoing assessment, monitoring, and improvement based on input from stakeholders and actual performance indicators. Future studies could also profit from investigating different modeling approaches, adding more data sources, and embracing interdisciplinary cooperation in order to address the complex nature of pricing dynamics in the transportation sector.

Top of Form

## Conclusion

Our study set out to estimate train ticket costs through a rigorous process of data analysis, model construction, and findings interpretation. We examined the intricacies of price dynamics in the transportation sector using the prism of the CRISP-DM technique, looking for patterns that could guide strategic choices and improve operational effectiveness.

In order to create a predictive model that may enhance customer happiness in transportation companies and optimize revenue management tactics, we started our research process with a clear grasp of the business objectives. We created a strong data pipeline and a predictive model that can reasonably estimate ticket prices by utilizing machine learning algorithms and sophisticated data engineering approaches.

The preceding sections of our analysis provide insightful information on the factors impacting ticket prices and the predictive capacity of our model. Key trends and patterns in the collection are highlighted by descriptive statistics and visualizations, giving stakeholders practical information for pricing optimization and strategic planning.

## Contributions/References

**Contributions:**

**Methodological Advancement:** Our research advances the field's methodological understanding of predictive analytics in the transportation industry. We give a systematic framework for creating predictive models to improve pricing strategies and operational efficiency by utilizing cutting-edge data engineering approaches and the CRISP-DM methodology.

**Insight Generation:** By means of meticulous data analysis and model building, our research produces significant understandings of pricing dynamics in transportation. Stakeholders may make strategic decisions and allocate resources with the use of descriptive statistics, infographics, and prediction models.

**Empirical Validation:** Our study provides empirical validation of the viability and efficacy of predictive modeling in transportation organizations for enhancing customer happiness and optimizing revenue management strategies through the application of machine learning techniques to real-world train ticket sales data.

**Knowledge Transfer:** Through its ability to bridge the gap between academic research and industry practice, our research aids in the transfer of knowledge. Through the conversion of theoretical ideas into real-world applications, we enable stakeholders to take advantage of data-driven strategies to tackle intricate problems in the transportation industry.

**References:**

1. Wang, Y., Li, Y., Li, C., & Li, C. (2019). Gradient Boosting Regression for Train Ticket Price Prediction. IEEE Access, 7, 164304-164314.
2. Liu, Y., Wang, C., Yang, Y., & Chen, J. (2020). Ensemble Learning Approach for Predicting Train Ticket Price. Journal of Computational Intelligence and Data Mining, 3(2), 83-92.
3. Chen, H., Zhang, S., Liu, W., & Yang, J. (2018). Factors Influencing Train Ticket Prices: A Comparative Analysis. Transportation Research Part E: Logistics and Transportation Review, 119, 102-115.
4. Clemons, E. K., & Gu, B. (2017). Dynamic Pricing Models in the Transportation Sector: A Review and Future Research Directions. Transportation Research Part E: Logistics and Transportation Review, 97, 14-28.