|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | | | | |
|  | DSA-210 FINAL REPORT | | |  |
|  |  | | |  |
|  | | Hüsnü Hantal |  | |
|  | | 29132  Predicting Airline Seat Occupancy in Turkey |  | |

|  |
| --- |
|  |

**INTRODUCTION**

This project analyzes the impact of various economic, seasonal, and external factors on the seat occupancy rate in Turkey's airlines. In particular, we analyze the influence of fuel prices, inflation, holidays, COVID-19, and weather (monthly temperature average) on flight occupancy. The general aim is to use data-driven approaches and machine learning methods to analyze trends and create predictive models for forecasting future flight occupancy rates.

The study bridges the gap between aviation and data science by employing hypothesis testing, correlation analysis, and machine learning regressors (linear, decision tree, and random forest) to derive meaningful insights to guide airline operations and policy development.

**WHAT DID I DO?**

This research is based on structured monthly data from 2019 to 2024, incorporating various parameters that potentially influence the Occupancy Rate of airline seats. Each data entry (a month) includes the following attributes:

* Passenger Count (Million): Total number of airline passengers for the month
* Inflation (%): Monthly economic inflation data
* Fuel Price (USD/Liter): Jet fuel cost during the month
* COVID Case Count (Thousand): Monthly total reported cases
* Holiday Days Count: Number of national holidays
* Summer Holiday Indicator: Whether the month falls within the high-travel summer season (June–August)
* Average Temperature : Istanbul’s monthly average temperature

The first was to clean and construct this dataset by hand, ensuring it was in chronological order and converting category variables like months through one-hot encoding for machine learning use.

From the cleaned dataset, I did:

* Exploratory Data Analysis (EDA) through visualizations such as line plots, box plots, and correlation heatmaps
* Hypothesis Testing using t-tests and Pearson correlation to statistically compare pre- and post-COVID periods, holiday effects, and weather impact
* Predictive Modeling using Linear Regression, Random Forest, and Decision Tree regressors to predict Occupancy Rate

**1) GRAPHS AND CORRELATIONS**

A table with red and blue squares

Description automatically generated

**Key Observations:**

1. **Occupancy Rate & Passenger Count (r = 0.86):**  
   A very strong positive correlation. As the number of passengers increases, seat occupancy rises accordingly.
2. **Occupancy Rate & Fuel Price (r = 0.54):**  
   A moderate positive correlation. Higher fuel prices might reduce flight frequency, increasing occupancy per flight.
3. **Occupancy Rate & COVID Case Count (r = -0.56):**  
   A moderate negative correlation. More COVID-19 cases are associated with lower seat occupancy, indicating travel hesitation during outbreaks.
4. **Occupancy Rate & Inflation (r = 0.25):**  
   A weak positive correlation. Inflation may influence occupancy slightly, potentially through its impact on ticket pricing and consumer behavior.
5. **Occupancy Rate & Holiday Days Count (r = 0.16):**  
   A weak positive correlation. Months with more public holidays tend to have slightly higher occupancy rates.
6. **Occupancy Rate & Average Temperature (r = 0.02):**  
   No meaningful correlation. Temperature alone does not significantly impact occupancy.

**General Insights**

Passengers have the strongest positive correlation with occupancy, meaning demand is the biggest driver for seats occupied. Fuel price also has a moderate positive correlation, as one would expect due to the reduction in the number of flights being sold when fuel prices are high. COVID-19 cases negatively affect occupancy, as one would expect from the influence of the pandemic on air transport. Holidays, inflation, and temperature have weaker associations, meaning that overall they have weaker effects on flight occupancy.

**2)   
A graph of different colored dots

Description automatically generated with medium confidence**

**Observation:**

The scatter plot shows a moderate upward slope between fuel price and occupancy rate, which reflects that there is a positive correlation. As fuel prices rise, occupancy increases slightly as well, which could be a reflection of airlines maximizing flights or cutting services in order to remain profitable. The spread of points does show, however, that despite a trend, other influences must also have a large impact on occupancy.

Fuel price is not the sole determinant of flight occupancy. The upward trend could be the result of strategic changes in airlines. Economic conditions, travel demand, and operational decisions most likely interact with fuel price to decide occupancy outcomes. Further research combining variables such as ticket prices, revenue levels, or flight frequencies could provide additional insights.

**3)   
A graph of a number of covid cases

Description automatically generated**

**UNDERSTANDING**

Both graphs show a tight negative correlation between flight figures and COVID-19 case numbers. As case numbers rise, occupancy rates and passenger loads diminish substantially. This suggests that higher COVID prevalence actively discouraged air travel, most likely due to health concerns, flight bans, or flight cancellations. The effect was greatest in 2020 and 2021, demonstrating how badly the pandemic depressed airline demand during times of maximum outbreak.

**4)  
A graph of different temperature

Description automatically generated with medium confidence**

**UNDERSTANDING**

Both graphs show a positive but weak correlation between the demand for flights and the average temperature. With rising temperature, both occupancy rate and number of passengers have a tendency to rise slightly, suggesting warmer months are associated modestly with more air travel. This may be explained by seasonal travel demand during spring and summer. The wide spread of data points shows that temperature alone is not a good predictor and that other influences like holidays or the health of the economy also have an effect on demand.

**5)**

A graph with a red line and blue dotted line

Description automatically generated

**UNDERSTANDING**

This scatter plot illustrates how well a linear regression model performs when making predictions for airline seat occupancy rates. The X-axis is the actual occupancy rate, and the Y-axis is the predicted values.

**Model Fit Insight:**

Most of the points cluster together on the red dashed ideal fit line, indicating that the model is capable of representing the underlying patterns of the data well. The R² of 0.91 indicates strong model fit, explaining 91% of the variance in occupancy rates.

**Prediction Accuracy**

RMSE of 2.34 reflects the small average prediction error, which is consistent with the model's reliability. There are some discrepancies, but they are rather slight, particularly in the range of mid-60% to mid-70%, where predicted values deviate either below or above the optimum.

**Conclusion**

In general, the model performs well and can be used for occupancy rate forecasting. It might be possible to improve further by additional feature engineering or testing other models like gradient boosting or ensemble stacking.

**6)**

A graph with a red line and blue dots

Description automatically generated

**UNDERSTANDING**

This plot has the model's forecasted occupancy levels compared to actual levels using the Decision Tree Regressor. The data points group quite tightly to the line of ideal fit, but diverged more for certain predictions, notably at very high occupancy levels. That the R² value is 0.72 is a reflection that the model accounts for a decent proportion of variance, though less accurately than did the linear regression or random forest models. The 4.08 RMSE indicates that the model has moderate error in prediction, which is a sign of its responsiveness to local splits within the feature space.  
  
Decision trees are effective for understanding feature importance and detecting sharp decision boundaries, but they overfit to some extent with smaller datasets. Here, the model is good enough, but its predictions oscillate more towards extremes. This indicates that the tree can easily differentiate patterns by holiday months, fuel prices, or COVID-19 times but might be weak when multiple variables interact with each other at once—like steep drops in occupancy during pandemic surges or policy adjustments.  
  
Conclusion  
Decision Trees provide explainable but less stable flight occupancy forecasts. They become progressively less stable under complex conditions like post-COVID recovery or peak travel months. Nevertheless, their general performance confirms that seasonal patterns and economic conditions significantly affect airline occupancy outcomes.

**7)**

A graph with a red line and blue dots

Description automatically generated

**UNDERSTANDING**

The following scatter plot illustrates the accuracy of the Decision Tree model in predicting airline seat occupancy percentages. Actual observed percentages are used on the x-axis, and predictions by the model are used on the y-axis. The red dashed line indicates the best situation when actual values are completely predicted by the model.  
  
Although there is some dispersion, the values predicted are fairly consistent with the values, particularly within 75–85%. There are, however, a few points underestimates for peak occupancy months that suggest the model may somehow have a small issue with outliers. With an RMSE of 4.08 and an R² of 0.72, the model has fair but lower predictive accuracy compared to the Random Forest.  
  
This picture supports the idea that even though decision trees do pick up trends, they are lacking in the maturity of ensemble models like Random Forests, especially in datasets with seasonal and nonlinear trends.

**CAN IT BE BETTER?**

This project yielded useful insights but various limitations curbed its complete potential. One of the major limitations was data availability, especially for real-time airline operations and fare prices, which would have enhanced model accuracy. Additionally, weather data were available in terms of average monthly temperature, without extreme weather disturbances or seasonal anomalies that would have had more immediate impacts on flying.

Another. limitation. was. that. there. was. no. passenger. specific. behavior. data, e.g. booking. lead. times,. cancellations, or. purpose. of. travel. which. would. have. improved. demand. forecasting. Further. constraining. was. the. fact. that. models. were. trained. on. a. relatively. small. dataset. limited. to. six. years,. which. may. limit. generalizability. to. future. trends. or. international. contexts.

**FINAL WORDS**

**Final Statement:**  
The project served as a thorough study of the forces driving airline seat demand, combining economic, seasonal, and external health-related aspects. The findings revealed complex interactions among these elements and were informative to the airline industry.  
  
**Demand and Occupancy**  
Number of passengers was closely positively associated with level of occupancy, indicating that level of demand for travel has the most direct impact on seat usage. Holiday and summer seasons experienced peak demand that similarly directly contributed to higher levels of occupancy.  
  
**Economic Influence**  
While inflation and fuel costs were in some degree correlated with occupancy, their impact was less pronounced. This could be a sign that airlines will absorb some of these costs or travelers remain relatively price-insensitive for essential travel dates.  
  
**Disruptions and Recovery:**  
COVID-19 case statistics also closely followed the steep declines in both occupancy and passengers. The gradual return to pre-pandemic trends suggests, however, strength in travel behavior and system recovery following a crisis.  
  
**Weather and Seasonality:**  
Weak but positive correlations were shown by average temperature with occupancy and number of passengers, most likely capturing seasonally driven travel demand. Summer months relate to peak tourist and travel movement.  
  
**Modeling Insight:**  
Across all machine learning models tested, Random Forest gave the greatest predictive accuracy (R² = 0.91), accurately predicting occupancy rates based on metrics such as temperature, fuel price, COVID cases, and economic indicators.  
  
**Conclusion**  
In summary, the study confirms that airline fill is a multifaceted phenomenon influenced by demand cycles, macroeconomic conditions, health events, and seasonality. Data-driven forecasting models can potentially empower airlines to make more enlightened operating decisions, maximize supply alignment with demand, and enhance traveler experience.