**Week 2**

**Table of Contents – Week2**

[Summary 1](#_Toc183182505)

[Milestones achieved 1](#_Toc183182506)

[Conclusions 2](#_Toc183182507)

[Data Analysis and Exploration 2](#_Toc183182508)

[Data Quality and Missing Values 2](#_Toc183182509)

[Feature Types and Encoding 3](#_Toc183182510)

[Feature Selection 3](#_Toc183182511)

[Key observations: 4](#_Toc183182512)

[Correlation 5](#_Toc183182513)

[References 6](#_Toc183182514)

[Challenges 7](#_Toc183182515)

[Next Steps 8](#_Toc183182516)

# Summary

This week, significant progress was made in the feature selection. The provided datasets were preprocessed to remove irrelevant columns, impute missing values, and encode features. Dimensionality reduction and correlation analysis guided feature selection for subsequent clustering and machine learning tasks.

# Milestones achieved

* Examined data through visualization and analysis techniques.
* Generated statistical summary
* Feature correlation to output class(es).
* Generated Scatter plot
* Dealing with Categorical data
* Fixed problems like missing values, errors or outliers. Missing values were imputed using statistical measures such as median and mode, depending on the context.
* Apply pre-processing or normalization procedures.
* Columns like ID, Customer Type and Satisfaction were removed. Costomer Type and Satisfaction are removed as Loyalty is derived from these.
* Categorical variables were encoded using LabelEncoder.
* Key features were selected based on Random Forest feature importance and backward elimination techniques.
* Interaction terms were created between numeric features to capture potential relationships that could improve prediction accuracy.

Conclusions

# Data Analysis and Exploration

Dataset 1 & 2: Dataset 1 & 2 are excel data as plotted below. The input CSV file contains 25 columns of customer data with variables capturing demographic details (e.g., Age, Gender), travel specifics (e.g., Flight Distance, Type of Travel), and service experience ratings across multiple aspects such as Online Boarding, Seat Comfort, and Food and Drink. The dataset also includes columns reflecting customer satisfaction and customer type from which Loyalty column is derived. This structured data is used to identify loyal customers.

## Data Quality and Missing Values

The dataset includes various customer attributes and satisfaction metrics aimed at predicting customer loyalty, segmented into four classes derived from **Customer Type** and **Satisfaction** ratings. To ensure data quality, an initial analysis was conducted to assess the completeness of each feature. While most features were complete, a few had minor data gaps. Specifically, **Arrival Delay** showed approximately 0.3% missing values, **Ease of Online Booking** had 4.37% missing values, and **Departure and Arrival Time Convenience** exhibited a 5.1% gap. Additionally, some in-flight service ratings, such as **Leg Room Service**, **In-flight Wifi Service**, and **Food and Drink**, had small data deficiencies.

A screenshot of a computer

Description automatically generatedA screenshot of a computer screen

Description automatically generatedA screenshot of a graph

Description automatically generated

Figure 1

A customized imputation strategy was employed to address these missing values effectively. For continuous features like **Arrival Delay** and **Departure Delay**, missing values were filled using the median, a robust choice against outliers. In cases where certain satisfaction metrics were recorded as zero due to missing feedback, these were replaced with the median of available responses, ensuring consistency in representation and reducing bias introduced by missing feedback.

## Feature Types and Encoding

The dataset contains both categorical and continuous variables, each requiring specific handling to prepare them for clustering and predictive modeling. Categorical features, including Gender, Customer Type, Type of Travel, and Class, were label-encoded to convert them into numerical representations that clustering algorithms can process. For example, Gender was transformed into binary values, while Type of Travel (Business or Personal) and Class (Business, Economy, Economy Plus) were similarly encoded to capture their distinct values numerically.

A group of blue dots

Description automatically generated A graph of a line

Description automatically generated with medium confidence

Continuous features, such as Flight Distance, Departure Delay, and Arrival Delay, were transformed to enhance their interpretability and reduce skewness. Binning was applied to the delay features based on quantiles to better capture the natural distribution of delays without excessive variance. Additionally, Age was binned into broader categories (e.g., young adults, middle-aged, seniors), simplifying age-related analysis and providing a more structured way to analyze demographic patterns.

## Feature Selection

After trying different approaches to select features, we concluded that relying on a single feature selection method (e.g., correlation) is insufficient for identifying the most impactful features. Instead, combining multiple feature selection techniques provides a more holistic view of feature importance. The binning of age and flight distance did not significantly enhance feature relevance, indicating the need for a more nuanced approach to numerical variables.

The table above combines ranks from various feature selection methods, including Correlation, KMeans ranking, Chi-square scores, Lasso, and Recursive Feature Elimination (RFE). By synthesizing these rankings, we identified the overall rank for each feature. This allows for the selection of the top 13 features, ensuring coverage across the top 10 features of individual ranking methods. This approach mitigates the risks of overfitting and underfitting by maintaining a balanced and robust set of features.

|  |  |  |  |
| --- | --- | --- | --- |
| **Class** | **Top 2 Features** | **Top 5 Features** | **Top 10 Features** |
| Class 1 | Online Boarding, Type of Travel | Online Boarding, Type of Travel, Age, Flight Distance, Departure and Arrival Time Convenience | Online Boarding, Type of Travel, Age, Flight Distance, Departure and Arrival Time Convenience, Seat Comfort, Ease of Online Booking, In-flight Wifi Service, In-flight Entertainment, On-board Service |
| Class 2 | In-flight Wifi Service, Ease of Online Booking | In-flight Wifi Service, Ease of Online Booking, Online Boarding, Type of Travel, Departure and Arrival Time Convenience | In-flight Wifi Service, Ease of Online Booking, Online Boarding, Type of Travel, Departure and Arrival Time Convenience, Age, Flight Distance, Baggage Handling, In-flight Service, Class |
| Class 3 | Type of Travel, Class | Type of Travel, Class, Online Boarding, In-flight Entertainment, Leg Room Service | Type of Travel, Class, Online Boarding, In-flight Entertainment, Leg Room Service, In-flight Wifi Service, Seat Comfort, On-board Service, Cleanliness, Baggage Handling |
| Class 4 | Online Boarding, In-flight Entertainment | Online Boarding, In-flight Entertainment, Class, Type of Travel, Seat Comfort | Online Boarding, In-flight Entertainment, Class, Type of Travel, Seat Comfort, Leg Room Service, Cleanliness, Flight Distance, On-board Service, In-flight Wifi Service |

## Key observations:

* **Type of Travel**, **Online Boarding**, and **In-flight Wifi Service** consistently rank high across multiple methods, indicating their strong predictive power.
* Features like **Age**, **Ease of Online Booking**, and **Flight Distance** also emerge as crucial, despite moderate rankings in some methods, due to their cumulative importance across categories.
* Lower-ranked features like **Departure Delay** and **Arrival Delay** show limited utility across ranking methods, suggesting they could be excluded without significant loss of predictive performance.

By focusing on the top-ranked 13 features, we can achieve a well-optimized feature set that balances comprehensiveness with simplicity, ultimately leading to improved model performance and generalizability.

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| **Feature** | **TC Rank** | **Correlation Rank** | **Kmean Rank** | **Chi2**  **Score** | **Lasso** | **RFE** | **Overall rank** |
| Type of Travel | 1 | 7 | 1 | 3 | 2 | 2 | 3 |
| Online Boarding | 2 | 4 | 2 | 7 | 1 | 6 | 4 |
| In-flight Wifi Service | 3 | 17 | 3 | 6 | 14 | 1 | 7 |
| Ease of Online Booking | 4 | 18 | 4 | 8 | 13 | 4 | 9 |
| Age | 7 | 3 | 8 | 2 | 4 | 18 | 7 |
| In-flight Entertainment | 8 | 6 | 9 | 11 | 5 | 7 | 8 |
| Flight Distance | 9 | 2 | 7 | 1 | 3 | 21 | 7 |
| Departure and Arrival Time Convenience | 6 | 16 | 6 | 10 | 12 | 9 | 10 |
| Seat Comfort | 11 | 5 | 10 | 12 | 7 | 15 | 10 |
| Class | 10 | 15 | 5 | 9 | 22 | 3 | 11 |
| Cleanliness | 15 | 8 | 14 | 14 | 17 | 12 | 13 |
| On-board Service | 13 | 11 | 11 | 13 | 9 | 11 | 11 |
| Leg Room Service | 14 | 10 | 13 | 15 | 8 | 14 | 12 |
| In-flight Service | 12 | 19 | 12 | 18 | 16 | 8 | 14 |
| Gate Location | 5 | 22 | 19 | 21 | 11 | 5 | 14 |
| Baggage Handling | 16 | 13 | 16 | 19 | 10 | 13 | 15 |
| Check-in Service | 17 | 14 | 15 | 16 | 22 | 10 | 16 |
| Food and Drink | 18 | 12 | 18 | 20 | 15 | 16 | 17 |
| Gender | 19 | 9 | 17 | 17 | 6 | 17 | 14 |
| Arrival Delay | 20 | 20 | 20 | 4 | 22 | 19 | 18 |
| Departure Delay | 21 | 21 |  | 5 | 22 | 20 | 18 |

# Correlation

* **Loyalty and Satisfaction**:

There is a **strong positive correlation** of **0.64** between **Loyalty** and **Satisfaction**, indicating that satisfied passengers are more likely to be loyal.

* **Satisfaction and Online Boarding**:

**Satisfaction** has a significant correlation of **0.50** with **Online Boarding**, suggesting that ease and convenience during boarding play a role in overall satisfaction.

* **Seat Comfort and Satisfaction**: **Seat Comfort** is correlated with **Satisfaction** at **0.35**, meaning that better seating contributes positively to passenger satisfaction.
* **Arrival Delay and Departure Delay**: These two metrics have the strongest correlation of **0.96**, reflecting that longer departure delays almost always lead to longer arrival delays. However, both have a very low correlation with **Satisfaction**, showing that delays, while frustrating, are not the biggest drivers of dissatisfaction.
* **In-flight Services**: **In-flight Entertainment** (0.62) and **Food and Drink** (0.62) are strongly correlated with one another, indicating that passengers often rate these services similarly. However, their correlation with **Satisfaction** is moderate, suggesting that while important, these services alone are not the main drivers of satisfaction.
* The strongest drivers of **Satisfaction** appear to be **Online Boarding**, **Seat Comfort**, and **In-flight Services**. The high correlation between **Loyalty** and **Satisfaction** emphasizes the importance of focusing on overall satisfaction to retain loyal passengers. Improving these key service areas may lead to increased satisfaction and, consequently, higher loyalty.
* Hence**,** based on the correlation values, we can decide which features to prioritize in our model. Features with higher positive correlations could be key predictors, while features with low or negative correlations may be less useful.
* Preprocessing was applied to deal with missing data as they can negatively affect model performance
* Different features in our dataset are on different scales (e.g., Age ranges from 18-60, while Flight Distance ranges from 0-5000+). Some machine learning algorithms (e.g., KNN, SVM, neural networks) are sensitive to the magnitude of values and will give more weight to features with larger numerical ranges.
* Hence, we normalize or standardize the numerical features, so they are on the same scale.
* The feature selection is done based on ranked output from different algorithms as mentioned below
  + Correlation Rank
  + KBest
  + LASSO Coefficient (Lasso from sklearn.linear\_model)
  + Chi-Square
  + Recursive Feature Elimination (RFE from from sklearn.feature\_selection)
  + ExtraTreesClassifier from sklearn.ensemble
* The algorithms highlight which features are most relevant to predicting the target loyalty value. For example, **Type of Travel** is ranked as the most important feature by the Extra Trees Classifier. This suggests that the nature of a customer’s travel (e.g., business or personal) can be a strong predictor of their loyalty type. While this feature may not exhibit a high correlation with the target variable in a traditional sense, it provides actionable insights into customer behavior, showcasing the value of ensemble-based feature selection methods like Extra Trees.
* Similarly, features like **Online Boarding** and **In-flight Wifi Service**, ranked highly by Chi-square and KMeans clustering, respectively, reveal how customer interactions with digital and onboard services are critical to determining loyalty. These features might not appear significant in linear models or standalone correlation analyses but gain importance when considering categorical relationships and clustering trends. For instance, **Online Boarding** was ranked second overall due to its consistent significance across methods, underscoring its value in predicting loyalty type.
* This example demonstrates the need to leverage multiple feature selection techniques. A single method, such as correlation, might overlook key predictors, especially those with complex, non-linear relationships to the target variable. For instance, features like **Age** and **Flight Distance** show importance through their contributions across various methods despite not always being in the top ranks individually.
* By combining insights from various methods—such as tree-based models, clustering, and statistical tests—we can achieve a more comprehensive understanding of feature relevance. This multi-method approach ensures the model incorporates diverse patterns in the data, leading to a better selection of features that collectively contribute to accurate predictions.

# References

* KDNuggets, 2021. *Feature Selection Overview: Selecting the Right Features for Your Model.* [online] Available at: <https://www.kdnuggets.com/2021/06/feature-selection-overview.html> [Accessed 21 November 2024].
* KDNuggets, 2020. *Getting Started with Feature Selection.* [online] Available at: <https://www.kdnuggets.com/2020/08/getting-started-feature-selection.html> [Accessed 21 November 2024].
* Towards Data Science, 2020. *Feature Selection Techniques in Machine Learning with Python.* [online] Medium. Available at: <https://towardsdatascience.com/feature-selection-techniques-in-machine-learning-with-python-f24e7da3f36e> [Accessed 21 November 2024].
* Scikit-learn, 2024. *sklearn.feature\_selection.SelectKBest.* [online] Available at: <https://scikit-learn.org/stable/modules/generated/sklearn.feature_selection.SelectKBest.html#sklearn.feature_selection.SelectKBest> [Accessed 21 November 2024].
* Kaggle, 2024. *Airline Passenger Satisfaction Dataset.* [online] Available at: <https://www.kaggle.com/datasets/teejmahal20/airline-passenger-satisfaction/data> [Accessed 21 November 2024].
* ProQuest Ebook Central, 2024. *Ebook Central Reader.* [online] Available at: <https://ebookcentral.proquest.com/lib/hw/reader.action?docID=30168989&ppg=67> [Accessed 21 November 2024].
* Géron, A., 2022. *Hands-On Machine Learning with Scikit-Learn, Keras, and TensorFlow.* O’Reilly Media, Incorporated, Sebastopol. Available from: ProQuest Ebook Central. [29 September 2024].
* GitHub, 2024. *End-to-End Machine Learning Project - Chapter 2.* [online] Available at: <https://github.com/ageron/handsonml2/blob/master/02_end_to_end_machine_learning_project.ipynb> [Accessed 21 November 2024].

# Challenges

* Struggling with concepts and understanding of data sets
* Understanding workflow
* Clusters weren’t formed correctly as the data set wasn’t good enough. We had to redo the feature selection with all the different algorithms.
* Identifying the right way to select features for complex data.
* Had challenge in terms of running the same set of values and deriving the selected features which made us to take up the extract and refer approach to extract the selected features under data folder “data/feature\_selected\_data.csv”.

# Next Steps

* Examining features by training classifiers and applying evaluation metrics
* Start studying the scikit-learn library of Python.
* To Search more about machine learning models to get some clarity about our upcoming task.
* Perform clustering of these selected features and validate if clustering is feasible.
* Extracted output to be clustered and model has to be trained.