**Week 2**

**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**

**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.

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Feature** | **Importance** | **TC Rank** | **Correlation Rank** | **Kmean Rank** | **Chi2**  **Score** | **Lasso** | **RFE** | **Overall rank** |
| Type of Travel | 0.157687 | 1 | 7 | 1 | 3 | 2 | 2 | 3 |
| Online Boarding | 0.114978 | 2 | 4 | 2 | 7 | 1 | 6 | 4 |
| In-flight Wifi Service | 0.109038 | 3 | 17 | 3 | 6 | 14 | 1 | 7 |
| Ease of Online Booking | 0.086492 | 4 | 18 | 4 | 8 | 13 | 4 | 9 |
| Age | 0.047013 | 7 | 3 | 8 | 2 | 4 | 18 | 7 |
| In-flight Entertainment | 0.046097 | 8 | 6 | 9 | 11 | 5 | 7 | 8 |
| Flight Distance | 0.04094 | 9 | 2 | 7 | 1 | 3 | 21 | 7 |
| Departure and Arrival Time Convenience | 0.052641 | 6 | 16 | 6 | 10 | 12 | 9 | 10 |
| Seat Comfort | 0.038821 | 11 | 5 | 10 | 12 | 7 | 15 | 10 |
| Class | 0.039614 | 10 | 15 | 5 | 9 | 22 | 3 | 11 |
| Cleanliness | 0.025059 | 15 | 8 | 14 | 14 | 17 | 12 | 13 |
| On-board Service | 0.029742 | 13 | 11 | 11 | 13 | 9 | 11 | 11 |
| Leg Room Service | 0.026262 | 14 | 10 | 13 | 15 | 8 | 14 | 12 |
| In-flight Service | 0.030306 | 12 | 19 | 12 | 18 | 16 | 8 | 14 |
| Gate Location | 0.057054 | 5 | 22 | 19 | 21 | 11 | 5 | 14 |
| Baggage Handling | 0.02316 | 16 | 13 | 16 | 19 | 10 | 13 | 15 |
| Check-in Service | 0.021555 | 17 | 14 | 15 | 16 | 22 | 10 | 16 |
| Food and Drink | 0.021267 | 18 | 12 | 18 | 20 | 15 | 16 | 17 |
| Gender | 0.011835 | 19 | 9 | 17 | 17 | 6 | 17 | 14 |
| Arrival Delay | 0.010298 | 20 | 20 | 20 | 4 | 22 | 19 | 18 |
| Departure Delay | 0.010139 | 21 | 21 |  | 5 | 22 | 20 | 18 |

**References**

* <https://www.kaggle.com/datasets/teejmahal20/airline-passenger-satisfaction/data>
* <https://ebookcentral.proquest.com/lib/hw/reader.action?docID=30168989&ppg=67>
* 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].
* <https://github.com/ageron/handsonml2/blob/master/02_end_to_end_machine_learning_project.ipynb>

**Challenges**

* Struggling with concepts and understanding of data sets
* Understanding workflow

**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.