Airline passenger satisfaction

Pravalika Battula *pxb220006*
Namratha Reddy Kareddy *nxk220058*
Haripriya Mutyala *hxm220050*
Paavana Sai Reddy Bandi *pxb220034*

***Abstract*—This report provides a predictive modelling of airline customer satisfaction using a number of supervised machine learning algorithms.The goal of this research is to understand the important aspects that determine a customer's degree of satisfaction with airline services. Along with this goal, this project is an attempt to comprehend the performance of various algorithms via the comparison of their result(s). Reasons for similarities and variances in these results are also offered. The conclusions taken from these results are explored in the concluding section.**

I. Introduction

Passenger satisfaction is a key performance indicator of the success of any airline company. Additionally, analyzing the elements that determine a passenger's degree of satisfaction would assist the airline firm in improving their Quality of Service, which might have an impact on their overall business growth.

The data set utilized in this study was taken from Kaggle and comprises of survey data from over 100,000 passengers. Around 25 factors are considered, such as "Type of Travel", "Class", Flight Distance, Ease of Booking, Check-in Service, and so on. Each passenger provides the values for each of these features. The final column, "Satisfaction," indicates whether the passenger was "satisfied" with the airline or felt "neutral/unsatisfied." As a result, this column serves as the label for all algorithms analyzed.

The dataset also has the following extra properties:

• Seven of the 25 features can accept continuous values. Such quantitative characteristics are standardized.

• Some passengers have missing values in columns such as "Arrival Delay." These null values are substituted with the column's mean.

• It is a nearly balanced data set in terms of the label under consideration.

• A 9:1 split of the dataset is used for training and testing.

II. DATA VISUALIZATION

Variable properties	Variable name		Flight operation quality	Departure arrival time convenient
Numerical type	Age	Satisfaction (0–5)	Ticketing service	Ease of online booking
	Flight distance			online boarding
	Departure delay in minutes		Ground service	Gate location
	Arrival delay in minutes			Baggage handling
Category type	Gender		Air service	Checking service
	Type of travel			Inflight Wi-Fi service
	Customer type			Food and drink
	Customer class			Seat comfort
				Inflight entertainment
				Onboard service
				Leg room service
				Inflight service
				Cleanliness

Fig. 1. Features

Above is a table (fig.1) which shows the data prescribed in the previous section.

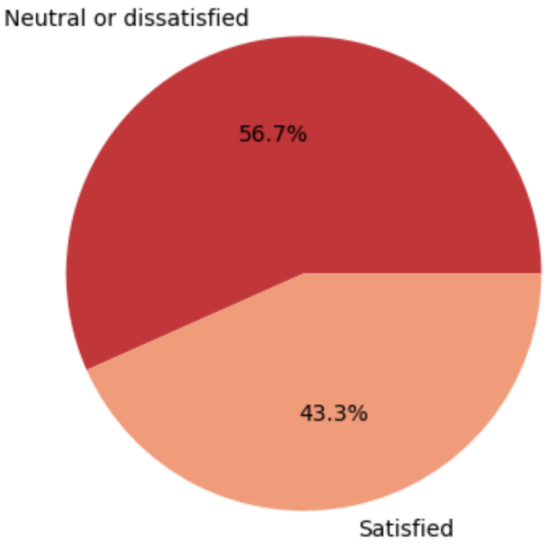

Fig. 2. Target variable

Examining the correlation matrix among features reveals a substantial correlation coefficient of 0.96 between "departure delay" and "arrival delay."

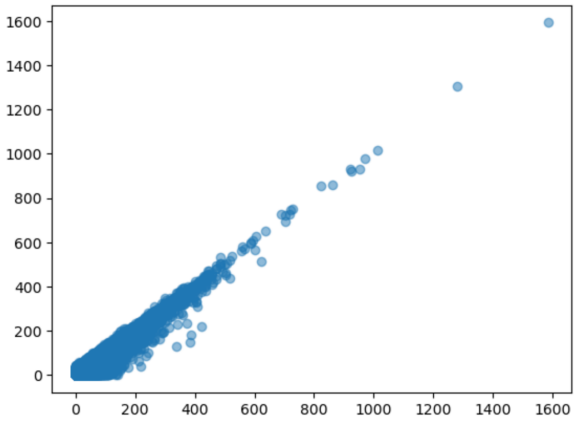

Fig. 3. Correlation between arrival and departure delays

The accompanying fig. 3 visually depicts the strong linear relationship between these two features. Consequently, to

streamline our models and mitigate multicollinearity, we opt to eliminate one of the features, namely "arrival delay," from the dataset.

III. Performance and Analysis of Algorithms

- Logistic Regression
- Decision Trees
- SVM (Support Vector Machine)
- KNN (K-Nearest Neighbours)

A. Logistic Regression

- Logistic regression is a statistical method used for binary classification, which predicts the probability of an instance belonging to a particular category. It's commonly employed in machine learning for tasks where the outcome is binary, such as determining whether an email is spam or not, or predicting whether a customer will make a purchase. Logistic regression uses the logistic function to model the relationship between the independent variables and the probability of the event occurring. Unlike linear regression, which predicts a continuous outcome, logistic regression outputs a probability score between 0 and 1, making it suitable for binary classification problems. The model is trained by adjusting its parameters to maximize the likelihood of the observed outcomes.
- Logistic regression employs the sigmoid function (also known as the logistic function) to transform a linear combination of input features into a value between 0 and 1.

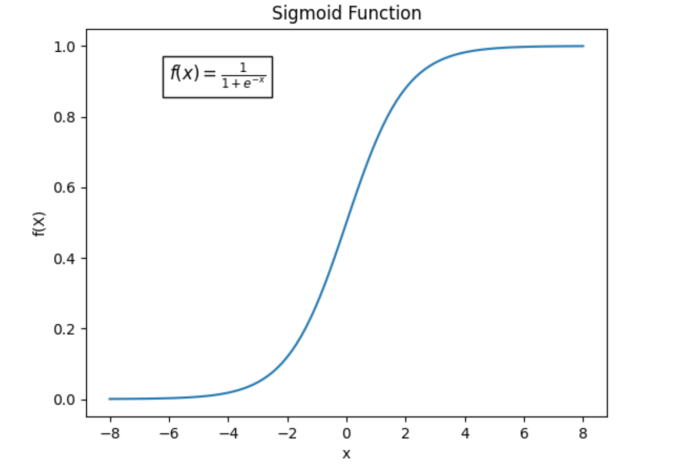

Fig. 4. Sigmoid Function

- Logistic regression is trained using a method called maximum likelihood estimation (MLE). The goal is to maximize the likelihood of observing the given set of outcomes under the logistic regression model. The cost function in logistic regression is the negative log-likelihood function. During training, the model's parameters are adjusted to minimize this cost function.
- Common evaluation metrics for logistic regression include accuracy, precision, recall, F1 score, and the area under the receiver operating characteristic (ROC) curve

When handling an extensive array of features, Logistic Regression, being a linear model, may encounter challenges in managing complexity. This complexity is reflected in the evaluation metrics derived from the test datasets. Notably, the ROC curve and its corresponding area under the curve highlight the difficulties arising from the elevated number of features in this particular scenario.

- AUC for the model:0.79

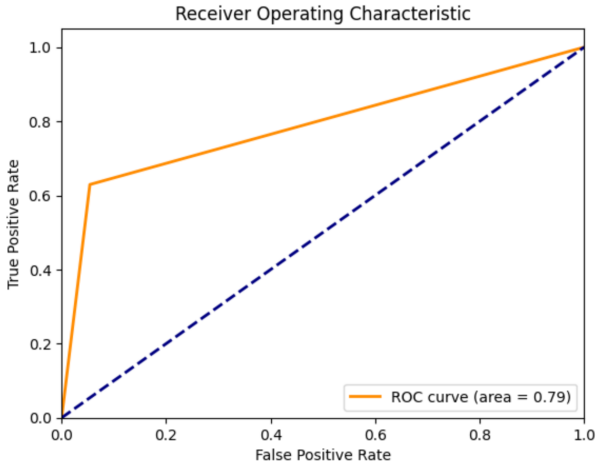

Fig. 5. ROC Curve for Logistoc Regression

Notably, the ROC curve (fig,4) and its corresponding area under the curve highlight the difficulties arising from the elevated number of features in this particular scenario.

The confusion matrix allows for a detailed assessment of the model's performance, and from these values, various metrics can be calculated, including:

- Accuracy: The overall correctness of the model, calculated as (TN + TP) / (TN + TP + FN + FP).
- Precision: The accuracy of the positive predictions, calculated as TP / (FP + TP). Precision provides insights into the model's ability to avoid false positive predictions.
- Recall (Sensitivity or True Positive Rate): The proportion of actual positives correctly predicted by the model, calculated as TP / (FP + FN). Recall indicates the model's ability to capture all positive instances.
- F1 Score: The harmonic mean of precision and recall, providing a balanced measure between the two metrics. It is calculated as 2 * (Precision * Recall) / (Precision + Recall).

To streamline the complexity of the model, focusing on the most crucial features is essential. This approach is aimed at enhancing the overall performance of the classifier. The report later provides details on how these significant features are identified, particularly in the sections discussing various classifiers.

	precision	recall	f1-score	support
0.0	0.77	0.95	0.85	5944
1.0	0.90	0.63	0.74	4447
accuracy			0.81	10391
macro avg	0.84	0.79	0.80	10391
weighted avg	0.83	0.81	0.80	10391

Fig. 6. Evaluation Matrix for Logistic Regression

B. Decision Trees

- A decision tree is a hierarchical structure composed of nodes, where each node represents a decision or a test on a feature.The top node is called the root, and the final nodes are called leaves, which contain the predicted output.
- Decision trees make decisions by recursively splitting the dataset based on the most informative feature at each node.The feature and split point are chosen to maximize the information gain (for classification) or reduce the mean squared error (for regression).
- For classification, decision trees often use metrics like information gain or Gini impurity to measure the effectiveness of a split.Information gain measures the reduction in uncertainty about the target variable after a split.
- Gini impurity measures the probability of misclassifying an instance chosen at random.A lower Gini impurity indicates a better split.

We implemented a Decision Tree model to predict airline passenger satisfaction. The initial iteration involved a shallow tree with a depth of 1, allowing for a simplified representation of decision boundaries. As we gradually increased the depth to 4, the tree became more complex, capturing finer details in the data. The decision tree structure evolved, revealing how different features contribute to predicting passenger satisfaction.

- AUC for the model:0.72(When D=4)

	precision	recall	f1-score	support
0.0	0.81	0.74	0.77	5944
1.0	0.69	0.76	0.72	4447
accuracy			0.75	10391
macro avg	0.75	0.75	0.75	10391
weighted avg	0.76	0.75	0.75	10391

Fig. 7. Evaluation matrix for Decision tree with D=1.

The decision tree model achieved an overall accuracy of 70 percent, indicating that 70 percent of the instances were correctly classified.Class imbalance was observed, with more instances of not satisfied passengers (Class 0) than satisfied passengers (Class 1).The model showed better performance in predicting satisfied passengers, with higher recall and F1-score for Class 1.Precision for not satisfied passengers (Class 0) was relatively high, but recall was lower, indicating that the model might miss a significant number of not satisfied passengers.

Classification Report:

	precision	recall	f1-score	support
0.0	0.85	0.57	0.68	5944
1.0	0.60	0.87	0.71	4447
accuracy			0.70	10391
macro avg	0.73	0.72	0.70	10391
weighted avg	0.75	0.70	0.69	10391

Fig. 8. Evaluation matrix for Decision trees with D=4.

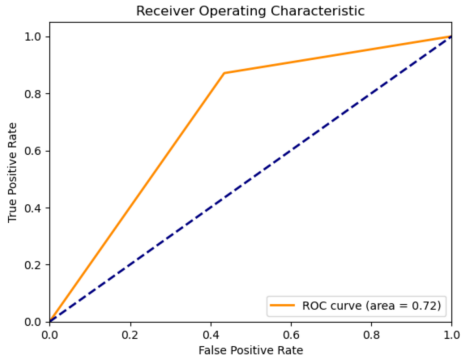

Fig. 9. ROc curve for Decision trees

C. SVM- Support Vector Machines

- Support Vector Machine (SVM) is a supervised machine learning algorithm used for both classification and regression tasks.
- SVM aims to find a hyperplane that best separates data points into different classes.For binary classification, the hyperplane should maximize the margin, which is the distance between the hyperplane and the nearest data points from each class.
- SVM can handle non-linearly separable data by transforming it into a higher-dimensional space using the kernel trick.Common kernel functions include linear, polynomial, radial basis function (RBF or Gaussian), and sigmoid.
- Support vectors are the data points that lie closest to the decision boundary (hyperplane).These points play a crucial role in defining the optimal hyperplane.
- The margin is the distance between the decision boundary and the nearest data point from each class.SVM aims to maximize this margin, providing a robust separation.
- SVM with non-linear kernels can create complex decision boundaries, allowing it to handle intricate patterns in the data.

The SVM was trained on a dataset comprising features (X) and corresponding labels (y), where labels were transformed to -1, 1 to facilitate binary classification. The iterative training process involved updating the model parameters, weight vector (w), and bias term (b), to optimize the classification boundary. The overall accuracy of the SVM model is approximately 39.25 percent.The confusion matrix indicates challenges in

```
Accuracy: 0.3925476603119584
Confusion Matrix:
 [[  0   0   0]
 [617   0  29]
 [ 55   0 453]]
Precision: 0.4137224303702798
Recall: 0.3925476603119584
F1 Score: 0.4028569928049997
```

Fig. 10. Evaluation results for Support Vector Machines

correctly predicting instances for all classes, especially for Class 0 where no instances are correctly predicted.The low precision, recall, and F1-score suggest that the model struggles to effectively classify instances into their respective classes. The calculated Area Under the Curve (AUC) serves as a

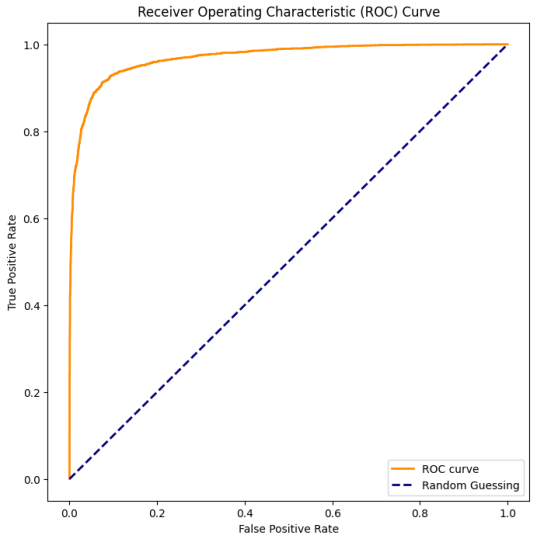

Fig. 11. ROc curve for Support Vector Machines

quantitative measure of the SVM's overall performance. A higher AUC value indicates improved discriminatory power, with 1.0 representing perfect classification.

- AUC for the model:0.96954

D. KNN-(K-Nearest Neighbour

- K-Nearest Neighbors (KNN) is a versatile and simple machine learning algorithm used for both classification and regression tasks
- KNN makes predictions based on the majority class (for classification) or the average value (for regression) of its k-nearest neighbors in the feature space.
- The hyperparameter 'k' represents the number of neighbors to consider when making predictions.A smaller value of k may lead to a more sensitive model, while a larger k may result in a smoother decision boundary.
- Common distance metrics include Euclidean distance, Manhattan distance, and Minkowski distance. The choice of distance metric depends on the nature of the data.
- KNN does not explicitly learn a decision boundary. Instead, it classifies a new data point based on the classes of its k-nearest neighbors.
- The choice of k influences the trade-off between bias and variance. Smaller k values may lead to a high-variance, low-bias model, while larger k values may result in a low-variance, high-bias model.

We employed the k-Nearest Neighbors (KNN) algorithm with a selected parameter k set to 5 to assess and predict passenger satisfaction in the airline dataset. This choice of k, after thorough experimentation with various values, demonstrated a balance between model complexity and predictive accuracy.

The KNN model with k=5 demonstrates strong performance with high precision, recall, and F1-scores for both classes.The overall accuracy of 91.16 percent indicates reliable classification across the dataset. The model appears to perform well in distinguishing between satisfied and not satisfied passengers, as evidenced by the balanced metrics in the classification report.

```
Classification Report:
              precision    recall  f1-score   support

         0.0       0.92      0.92      0.92       646
         1.0       0.90      0.90      0.90       508

    accuracy                           0.91      1154
   macro avg       0.91      0.91      0.91      1154
weighted avg       0.91      0.91      0.91      1154
```

Fig. 12. Evaluation Matrix for KNN

IV. Results and Analysis

Below are the results of the 4 algorithms

Algorithm	Accuracy
Logistic Regression	80%
Decision Trees (D=4)	70%
SVM	39%
KNN	91%

Fig. 13. Algorithm and their accuracy's

- on the following results, KNN seems to be the best model for the airline passenger satisfaction dataset with an accuracy of 91 percent.
- Logistic regression gave Good accuracy of 80 percent and is Well-suited for linear relationships in the data.
- Decision trees gave Limited performance with a reported accuracy of 70

Increasing the tree depth might improve performance but could lead to overfitting. SVM Reported the lowest accuracy among the models and is clear that it is not well suited for this dataset.

V. Conclusion

- Various algorithms, encompassing both linear and non-linear models, were examined and scrutinized in terms of their performance. Initially, linear models such as Logistic Regression exhibited lower accuracy values, likely attributed to a high number of features.
- o address this, we enhanced accuracy by selectively incorporating only the most influential features. These critical features were identified through the outcomes of a decision tree classifier. The process of feature selection significantly contributed to refining the model's accuracy.
- The optimization of algorithmic hyperparameters was carried out through multiple iterations, exploring different values for these parameters. The results of these iterations were meticulously plotted and illustrated in corresponding figures. This exploration not only aids in pinpointing the optimal hyper-parameter values but also provides insights into instances of underfitting or overfitting, contributing to a nuanced understanding of model behavior.
- After a thorough analysis, the best hyperparameter values were carefully selected. Subsequently, a comprehensive evaluation of the model's performance was conducted, encompassing various metrics. These metrics serve as valuable indicators of the model's efficacy and help draw informed conclusions about its generalization capabilities.

VI. Future Work

- The current set of features might not encompass all the factors that influence passenger satisfaction. For instance, services like baggage transfer, which were not included in the initial feature set, could play a crucial role in shaping overall satisfaction.
- Advanced machine learning techniques, particularly neural networks, could be employed to explore and model intricate relationships within the data. Neural networks are adept at capturing non-linear patterns and might yield more accurate predictions by learning complex representations of the data.
- Integrate relevant external data sources to enrich the dataset and provide a more comprehensive view of the external influences on passenger satisfaction.
- Analyze trends and patterns that emerge over the long term, helping airlines anticipate shifts in passenger preferences and improve satisfaction strategies.

References

[1] Building Logistic Model - https://medium.com/towards-data-science/building-a-logistic-regression-in-python-step-by-step-becd4d56c9c8

[2] Decision Trees - https://christophm.github.io/interpretable-ml-book/tree.html

[3] Understanding how SVM Works - https://towardsdatascience.com/support-vector-machine-introduction-to-machine-learning-algorithms-934a444fca47

[4] Evaluation Metrics - https://www.analyticsvidhya.com/blog/2021/05/know-the-best-evaluation-metrics-for-your-regression-model/

[5] Accuracy Class and Metrics -https://keras.io/api/metrics/accuracy_metrics/accuracy-class