Predicting Flight Cancellation: A Machine Learning Approach

Smitha Kannur Ashok, Vishnu Vardhan Manivannan, Lokesh Bhushan, Lasvitha Pregada,
Sneha Venkatapathy, Sai Rohith Yadav Kalasani

Abstract:

The aviation industry operates within a dynamic environment where the reliability of flights is critical for the seamless functioning of the entire ecosystem. Flight cancellations, stemming from various factors, pose significant challenges for passengers and airlines alike. In this research, we employ machine learning models to predict flight cancellations, drawing insights from an extensive dataset from 2009 to 2018. Through a meticulous exploration of historical data, we aim to identify the key contributing factors, enhance predictive accuracy, and provide actionable insights for proactive decision-making in the aviation industry. This research aims to develop a comprehensive understanding of the factors that contribute to flight cancellations and identify opportunities for proactive decision-making to mitigate their impact. By analyzing historical data, we aim to develop accurate predictive models that can be used to anticipate cancellations and inform operational decision-making. Our research will contribute to the development of a more reliable and efficient aviation industry by providing airlines with the tools and insights necessary to make proactive decisions that minimize the impact of flight cancellations on their operations and passengers.

Keywords: *aviation industry, flight cancellation, machine learning, decision making.*

Introduction:

In the contemporary air travel landscape, the efficient management of flight operations is a critical concern. Flight cancellations, in addition to causing inconvenience to passengers, can result in operational and financial challenges for airlines. By utilizing machine learning techniques to predict flight cancellations, the proactive management of air traffic can be revolutionized, thereby minimizing disruptions and optimizing operational efficiency. This study delves into the multifaceted aspects of flight cancellations, aiming to develop a robust predictive model based on historical data. The proposed model is expected to enhance the ability of airline operators to manage their flight operations effectively and efficiently, ultimately leading to improved customer satisfaction and financial performance.

Problem Statement:

The primary aim of this study is to develop a reliable machine-learning model that can accurately predict flight cancellations. Through the analysis of past data on airline operations, weather conditions, and other relevant factors, we intend to create a predictive tool that can be used to make informed decisions. The goal is to minimize disruptions and improve overall operational efficiency by providing airlines with actionable insights.

Key Questions:

1. What are the primary factors contributing to flight cancellations?
   - To address this question, our research delves into the intricate relationships between various variables, aiming to identify and understand the key factors that contribute to flight cancellations.
2. Can machine learning models effectively predict flight cancellations with a high degree of accuracy?
   - The research focuses on evaluating the performance of machine learning models in predicting flight cancellations, aiming for a high level of accuracy and reliability in forecasting.
3. How can the insights derived from the predictive model be utilized to enhance airline operations and minimize cancellations?
   - This question directs our attention towards the practical implications of the predictive model. We aim to provide actionable insights that can be employed by airlines to optimize their operations and reduce the likelihood of cancellations.

Data Source:

The present study is based on a dataset obtained from Kaggle, a reputable data science platform that offers datasets and competitions. The dataset in question, entitled "Airline Delay and Cancellation Data,"[1] provides a comprehensive overview of air travel spanning over a decade. It comprises detailed information on flight delays and cancellations, including the root causes of disruptions, airport data, and weather conditions. This dataset is a valuable resource for researchers and analysts who wish to explore air travel patterns and the factors that impact their reliability.

The dataset contains the following columns:

i) FL_DATE: Date of flight

ii) OP_CARRIER: Airline Identifier

iii) OP_CARRIER_FL_NUM: Flight Number

iv) ORIGIN: Place of origin of flight

v) DEST: Flight Destination

vi) CRS_DEP_TIME: Planned Departure Time

vii) CRS_ARR_TIME: Planned Arrival Time

viii) CANCELLED: Whether the flight is canceled or not

ix) CANCELLATION_CODE: Reason for cancellation (A - Airline/Carrier; B - Weather; C - National Air System; D – Security)

x) DIVERTED: Whether the flight has been diverted

xi) CRS_ELAPSED_TIME: Planned time amount needed for the flight trip.

xii) Distance: Distance between origin and destination airports.

Data Pre-processing:

Data preprocessing is essential as it significantly improves the quality of data, leading to more accurate and efficient analytical outcomes. By cleaning, transforming, and normalizing data, preprocessing ensures that datasets are free of errors and inconsistencies, enabling models to learn from relevant, high-quality data. This process lays the foundation for effective data mining, predictive modeling, and decision-making, and is critical for extracting meaningful insights and achieving reliable results.

A. Data Cleaning:

- Size and Dimensions of the Data:

This project uses two years' worth of airflight data from 2017 and 2018. The memory size consumed by the data frame upon importing from the Comma Separated Values (CSV) files was 2.5GB. It contains 12,888,067 rows and 28 columns of data.

- Validating Data Types:

The data types of all the variables are as follows:

FL_DATE	OP_CARRIER	OP_CARRIER_FL_NUM	ORIGIN	DEST
"character"	"character"	"integer"	"character"	"character"
CRS_DEP_TIME	DEP_TIME	DEP_DELAY	TAXI_OUT	WHEELS_OFF
"integer"	"numeric"	"numeric"	"numeric"	"numeric"
WHEELS_ON	TAXI_IN	CRS_ARR_TIME	ARR_TIME	ARR_DELAY
"numeric"	"numeric"	"integer"	"numeric"	"numeric"
CANCELLED	CANCELLATION_CODE	DIVERTED	CRS_ELAPSED_TIME	ACTUAL_ELAPSED_TIME
"numeric"	"character"	"numeric"	"numeric"	"numeric"
AIR_TIME	DISTANCE	CARRIER_DELAY	WEATHER_DELAY	NAS_DELAY
"numeric"	"numeric"	"numeric"	"numeric"	"numeric"
SECURITY_DELAY	LATE_AIRCRAFT_DELAY	Unnamed..27		
"numeric"	"numeric"	"logical"		

Since FL_DATE is a date, we convert it from character format to date format.

- Removing Missing values:

Data with missing values can undermine the performance of statistical and machine learning models because these models require complete and consistent input data to accurately learn patterns and make predictions. Missing values can introduce bias, reduce the model's ability to generalize, and may lead to errors or misinterpretations in the results, compromising the reliability and effectiveness of the analysis. Show below are the variables in the dataset that contain missing values along with how many rows of missing values are present in that variable. Our approach to tackling this issue is to remove the columns that contain a high percentage of missing values. For the variables with a low percentage of missing values, we use a data imputation approach to fill in the missing values.

```
> print(nan_counts)
```

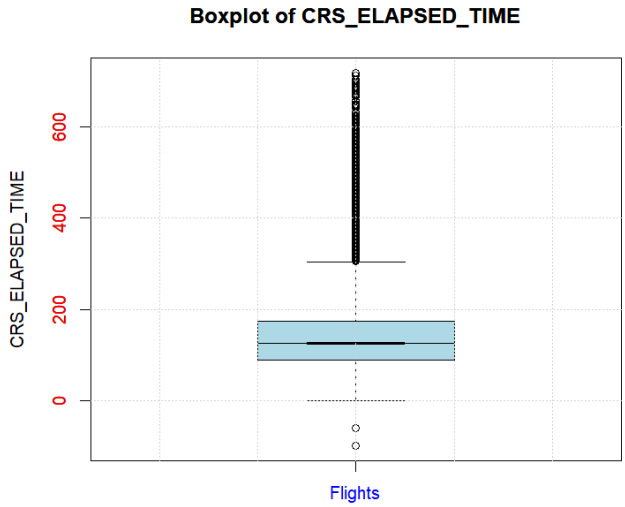
FL_DATE	OP_CARRIER	OP_CARRIER_FL_NUM	ORIGIN	DEST
0	0	0	0	0
CRS_DEP_TIME	DEP_TIME	DEP_DELAY	TAXI_OUT	WHEELS_OFF
0	192625	197577	197975	197970
WHEELS_ON	TAXI_IN	CRS_ARR_TIME	ARR_TIME	ARR_DELAY
203920	203920	0	203919	232251
CANCELLED	CANCELLATION_CODE	DIVERTED	CRS_ELAPSED_TIME	ACTUAL_ELAPSED_TIME
0	0	0	17	229653
AIR_TIME	DISTANCE	CARRIER_DELAY	WEATHER_DELAY	NAS_DELAY
229653	0	10505884	10505884	10505884
SECURITY_DELAY	LATE_AIRCRAFT_DELAY	Unnamed..27		
10505884	10505884	12888067		

These are the variables left after dropping the columns with high percentage of missing values:

```
 [1] "FL_DATE"           "OP_CARRIER"        "OP_CARRIER_FL_NUM" "ORIGIN"             "DEST"
 [6] "CRS_DEP_TIME"      "CRS_ARR_TIME"      "CANCELLED"         "CANCELLATION_CODE" "DIVERTED"
[11] "CRS_ELAPSED_TIME"  "DISTANCE"
```

- Data Imputation:

The variable CRS_ELAPSED_TIME contained only 17 missing values out of 12.8M rows. Due to this, the column has not been dropped. Instead, the missing values will be imputed. Several approaches can be used to impute this data. Upon inspection of the variable, we find that it has a large percentage of outliers.



Median is preferred over the mean for data imputation in the presence of many outliers because it is more robust and less sensitive to extreme values. Unlike mean, which can be significantly skewed by outliers, the median represents the middle value of a dataset, making it a better representation of the central tendency in skewed distributions. This ensures a more accurate and representative imputation of missing values, especially in datasets where outliers might otherwise distort the overall data analysis.

- Re-checking Size and Dimensions of the Data:

After processing the data, we find that the resultant data frame contains 12,888,067 rows and 12 columns of data. The memory consumption of the data frame is 1GB.

B. Exploratory Data Analysis & Feature Engineering:

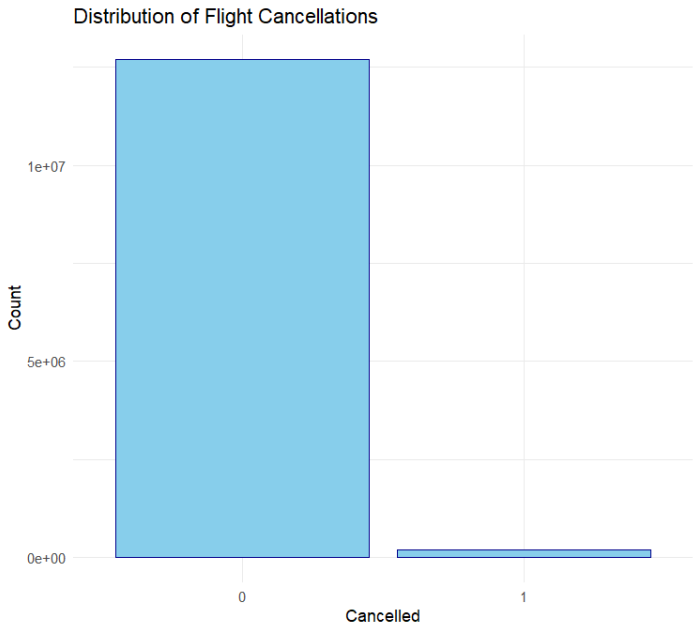
Exploratory Data Analysis (EDA) is a crucial step, where the primary objective is to explore and understand the data's underlying structure and characteristics. It involves summarizing the main features of a dataset, using visual methods like graphs and charts, and statistical techniques. EDA helps in identifying patterns, detecting anomalies or outliers, understanding data distribution, and exploring relationships between variables.

Feature engineering is the process of transforming raw data into meaningful features that significantly enhance the performance of machine learning models. It involves creating new features, selecting

relevant ones, and transforming existing variables to better capture the underlying patterns in the data, thereby improving model accuracy and efficacy. This step is crucial as the right features can greatly influence the success of predictive models.

- Flight Cancellations Distribution (target variable):

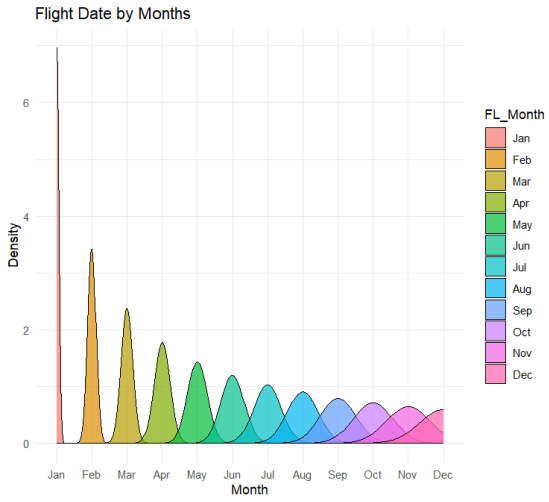
In the bar plot below, we observe that the target variable is highly skewed. Only a small proportion of flights get cancelled.



- Flight Date (Feature Engineering and Analysis):

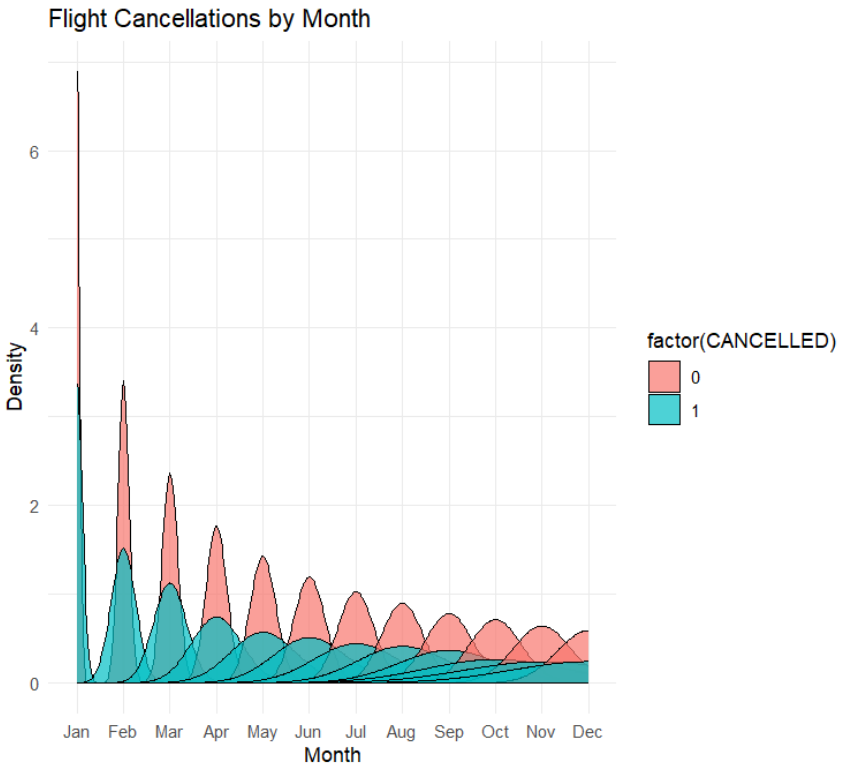
First, the flight month is extracted from the flight date and a new variable is created from it.

Frequency of flights for each month:

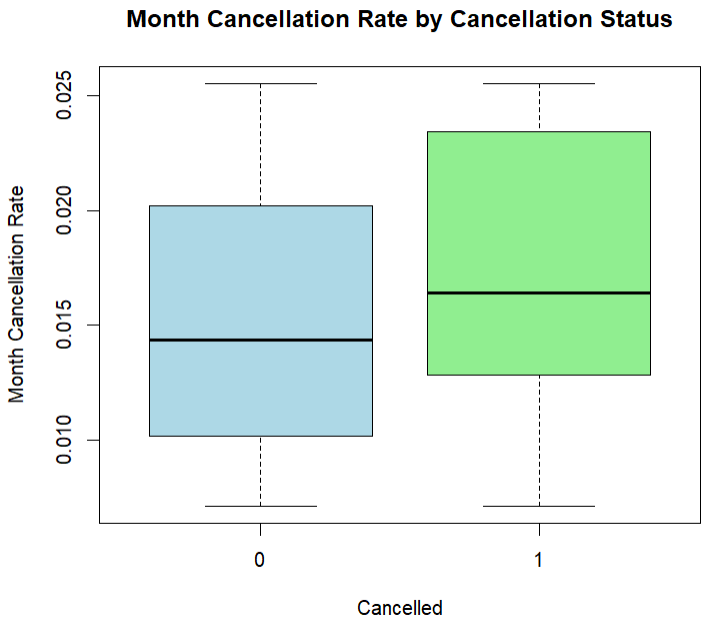


It is observed that the frequency of flights is highest in January, and it decreases over the months up to December.

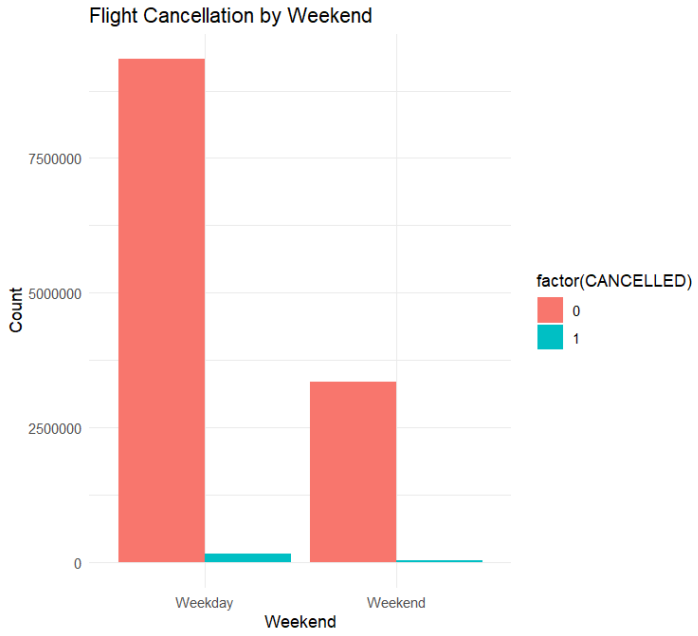
Flight cancellations per month:



A new variable is created for the flight cancellation rate per month. This is plotted against the target variable:

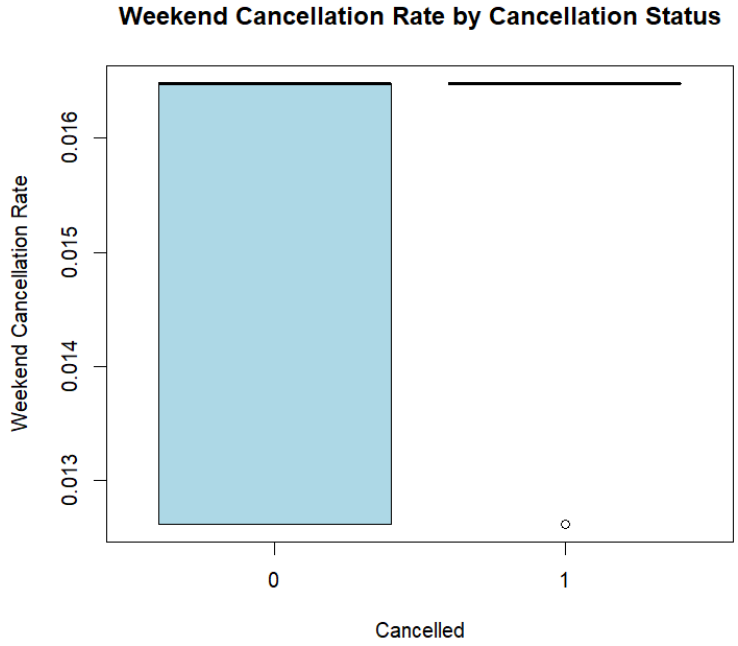


A new variable is created that classifies the date as weekend or weekday. This is plotted against the target variable.



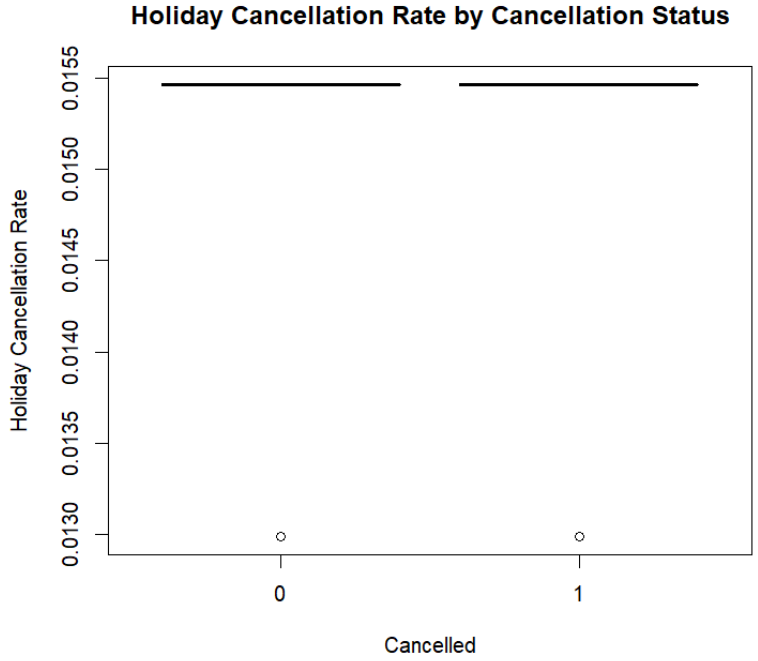
A significant difference in cancellations between weekday and weekend is not observed.

A new variable is created for the weekend cancellation rate, and this is plotted against the target variable.



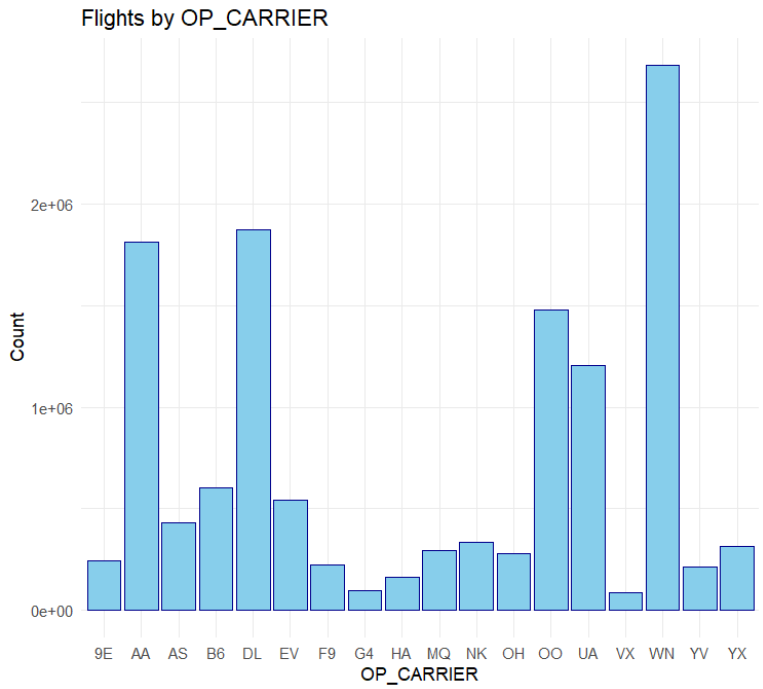
A new variable is created for whether the date is a holiday or not. This is the plotted against the target variable.

A new variable is created for the cancellation rate of flights during holidays. This is then plotted against the target variable.

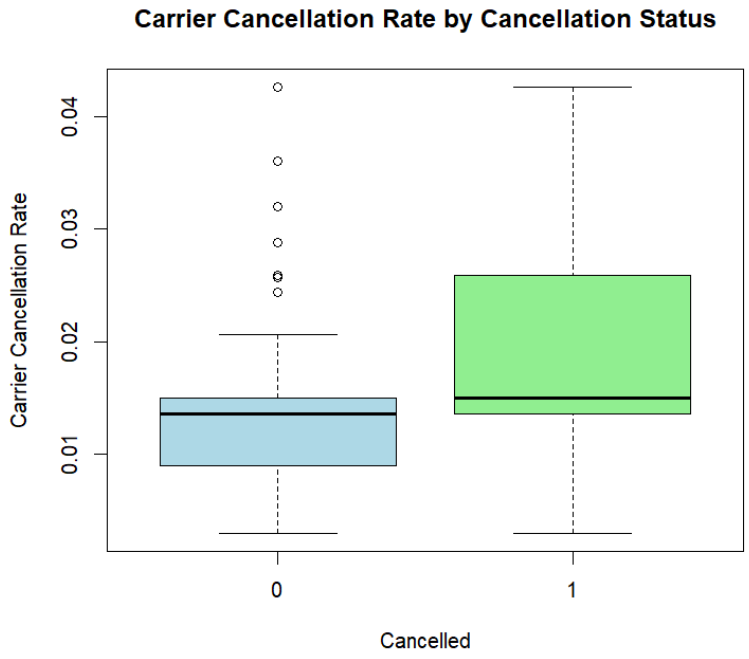


- Flight Carrier (Feature Engineering & Analysis):

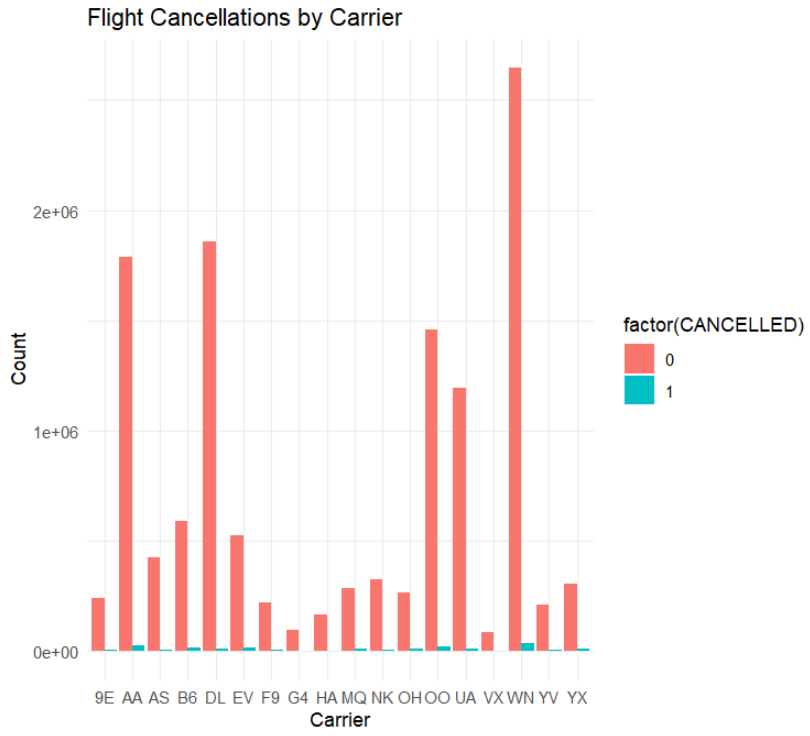
Distribution of flights by carrier:



A new variable is created for the flight cancellation rate by carrier. This is then plotted against the target variable:

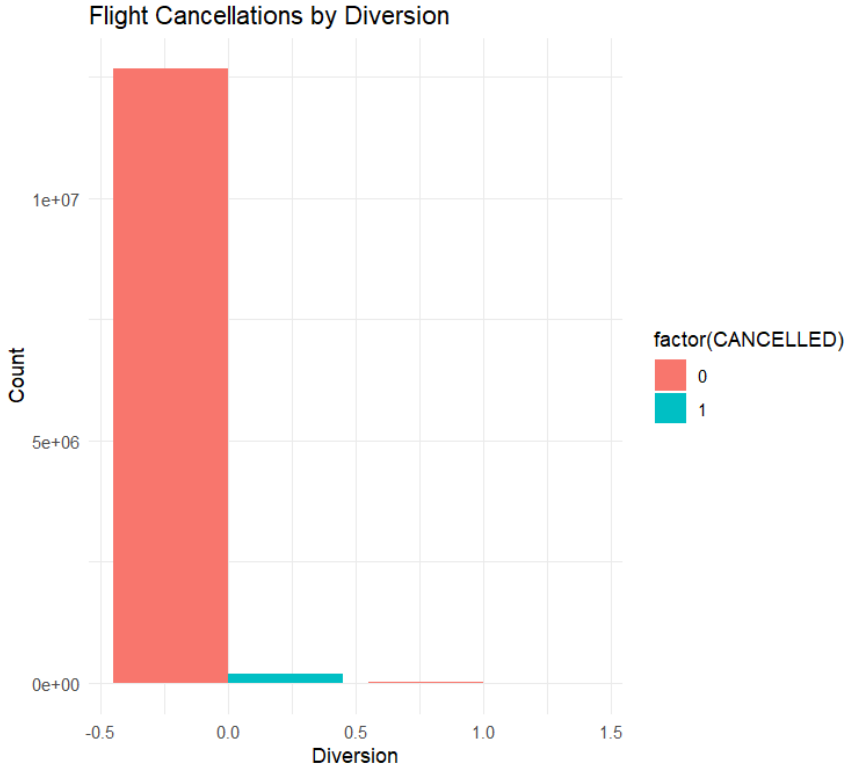


Distribution of Flight Cancelled/Not Cancelled by Carrier:

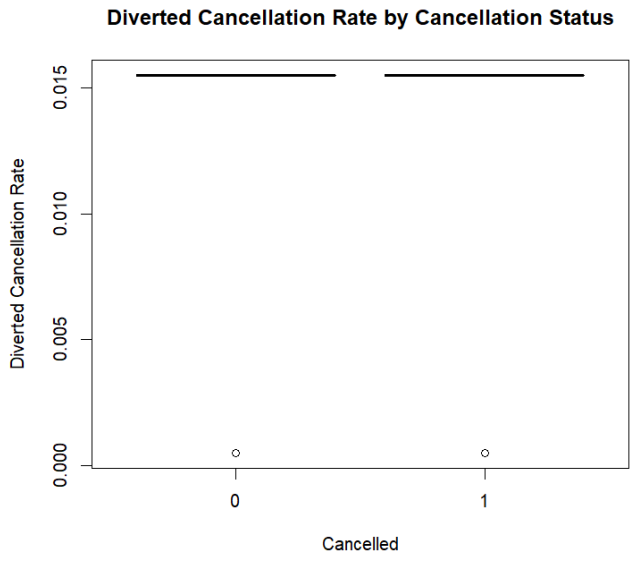


- Flight Diversions (Feature Engineering and Analysis):

Flight Cancellations by Diversion:

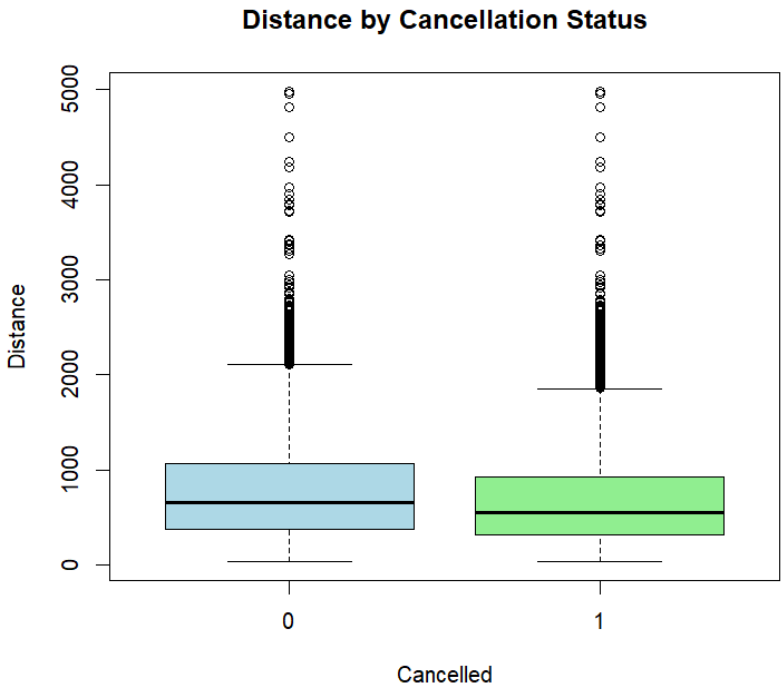


A new variable is created for the cancellation rate by diversions. This variable is plotted against the target variable:



- Flight Distance:

Here, the flight distance is plotted against the target variable:



C. Dropping Variables:

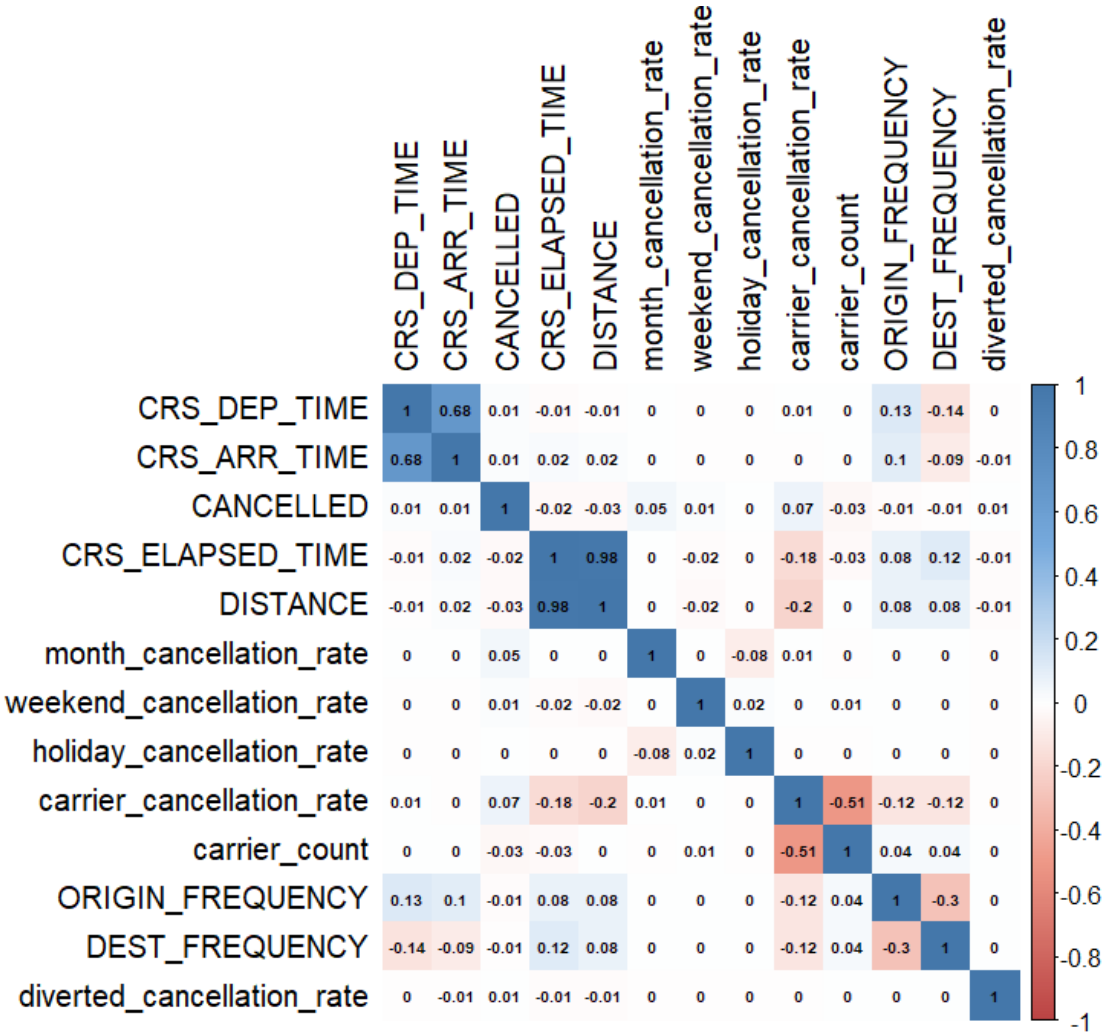
The features that do not show a meaningful relationship with the target variable as shown in the plots are dropped. Furthermore, the features from which other meaningful features have been extracted are also dropped as they are now redundant and will likely pose a multicollinearity problem with the extracted features.

Here are the final column names:

```
 [1] "CRS_DEP_TIME"              "CRS_ARR_TIME"              "CANCELLED"
 [4] "CRS_ELAPSED_TIME"          "DISTANCE"                  "month_cancellation_rate"
 [7] "weekend_cancellation_rate" "holiday_cancellation_rate" "carrier_cancellation_rate"
[10] "carrier_count"             "ORIGIN_FREQUENCY"          "DEST_FREQUENCY"
[13] "diverted_cancellation_rate"
```

D. Assessing Correlations:

A correlation matrix is a table showing correlation coefficients between variables, used to understand the strength and direction of relationships in a dataset. Each cell in the matrix represents the correlation between two variables, helping to identify potential connections and dependencies. It is very useful for feature selection and detecting multicollinearity.



It is observed that flight distance (DISTANCE) and the planned travel time (CRS_ELAPSED_TIME) are correlated with each other so one of them needs to be dropped. Flight distance is dropped as it is less correlated with flight cancellations (target variable) than the planned travel time. Holiday cancellation rate is also dropped as it has 0 correlation with the target variable.

Model Building:

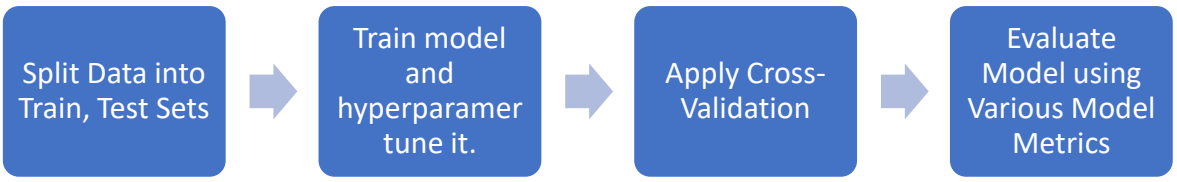
A. Methodology:

First, we split the data into training and testing sets. Then, we choose a statistical machine learning model of our preference. This model is then trained on the training set with different sets of hyperparameters to find the best possible combination of hyperparameters that result in the highest accuracy of the model.

Hyperparameter tuning is the process of optimizing the parameters that govern the learning process of a machine learning model to enhance its performance. This involves systematically searching for the ideal values of hyperparameters (such as the number of trees in a random forest or the number of neighbors in KNN), often using techniques like grid search or random search, to achieve the best possible accuracy or efficiency on a given dataset.

Furthermore, cross-validation is applied on the hyperparameter tuned model to assess its generalizability and robustness by evaluating its performance on different subsets of the data, in turn reducing the risk of overfitting. The model is then evaluated based on its accuracy and recall on testing data.

Recall, also known as sensitivity or true positive rate, is a metric that measures a model's ability to identify positive instances correctly out of all actual positive instances. In the context of flight cancellation prediction, recall is especially useful because it focuses on minimizing false negatives. In this scenario, a false negative would be predicting that a flight will not be cancelled when it is actually cancelled.



B. Splitting the Data:

The data is split into training and testing sets. 80% of the data becomes the training set and 20% becomes the testing set. The majority of the data is used in the training set to provide a comprehensive learning foundation for the model, ensuring it effectively learns and generalizes from a broad and representative sample of the data.

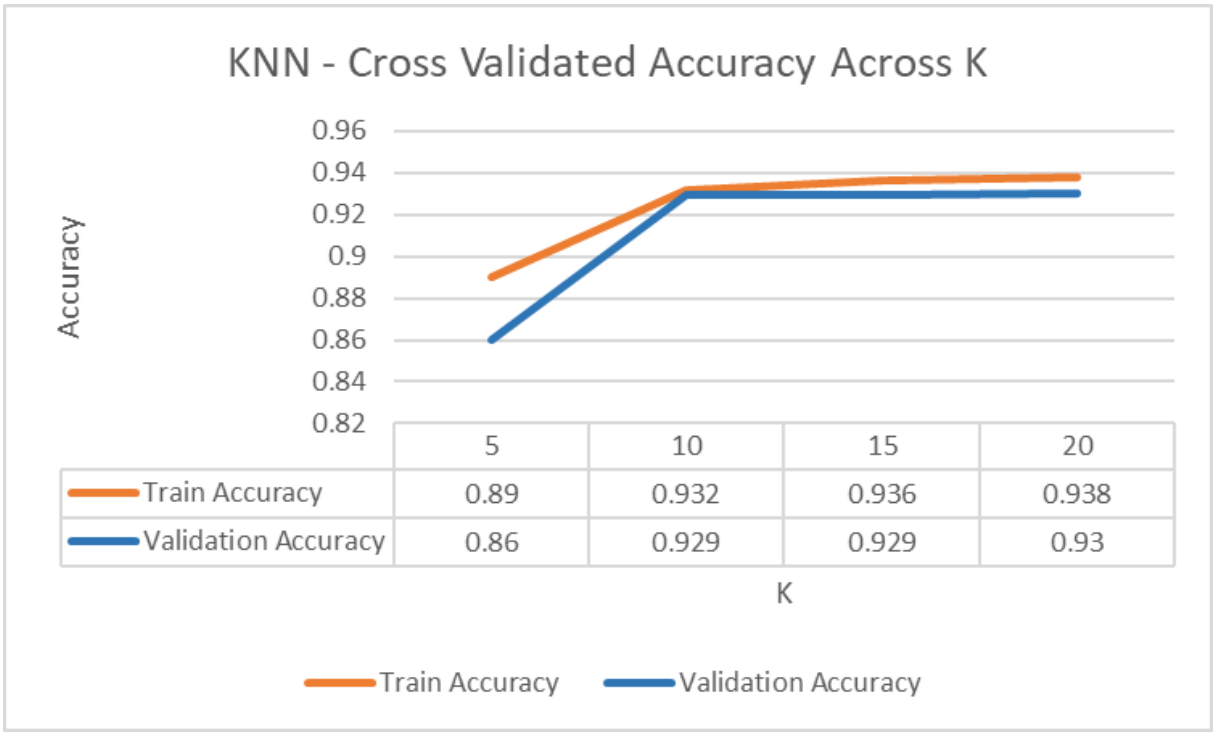
C. K-Nearest Neighbors:

K-Nearest Neighbors (KNN)[2] is a versatile, instance-based learning algorithm used for classification and regression. It functions by identifying the 'k' nearest data points in the feature space to a query point and bases its prediction on the aggregate output (majority class for classification, mean for regression) of these neighbors. Notable for its simplicity and effectiveness, KNN's performance is contingent on the choice of 'k' and the distance metric, though it may face challenges with large-scale data and feature relevance.

Hyperparameter Tuning: K is the only hyperparameter for the KNN model. Values between 5 to 20 at intervals of 5 are tested.

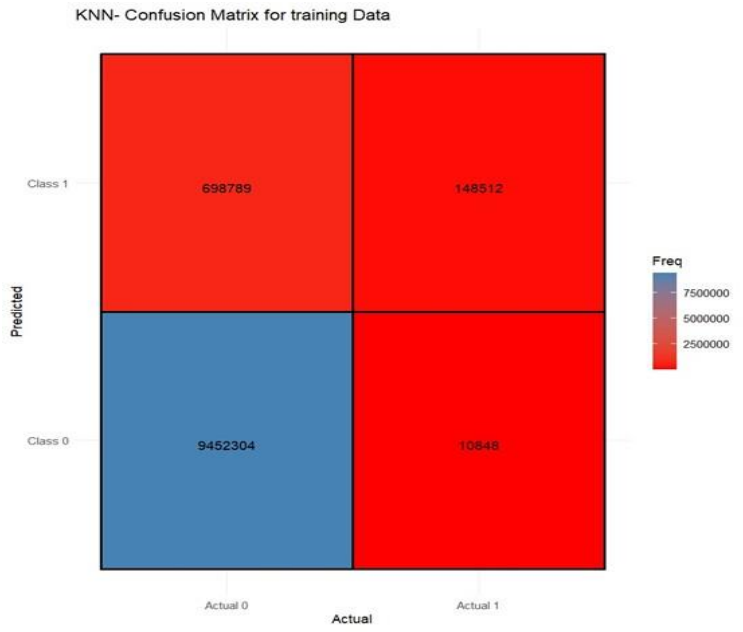
Cross-validation: 5 fold cross-validation is performed.

The training and validation accuracies for different values of K are as follows:

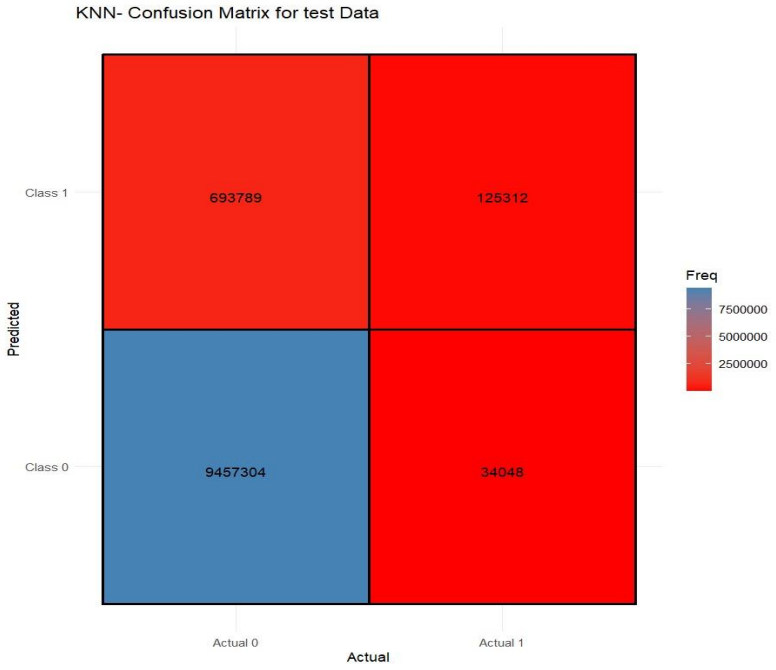


There's a significant increase in training and validation accuracies between K=5 and K=10 yet increasing K beyond 10 does not significantly increase the training and validation accuracy but increases the computational complexity. Therefore, we choose 10 as our final value for the hyperparameter K.

Confusion Matrix on training data (KNN):



Confusion Matrix for testing data (KNN):



KNN Evaluation Metrics		
	Accuracy	Recall
Training Set	93.11%	93.11
Testing Set	92.94%	93.16

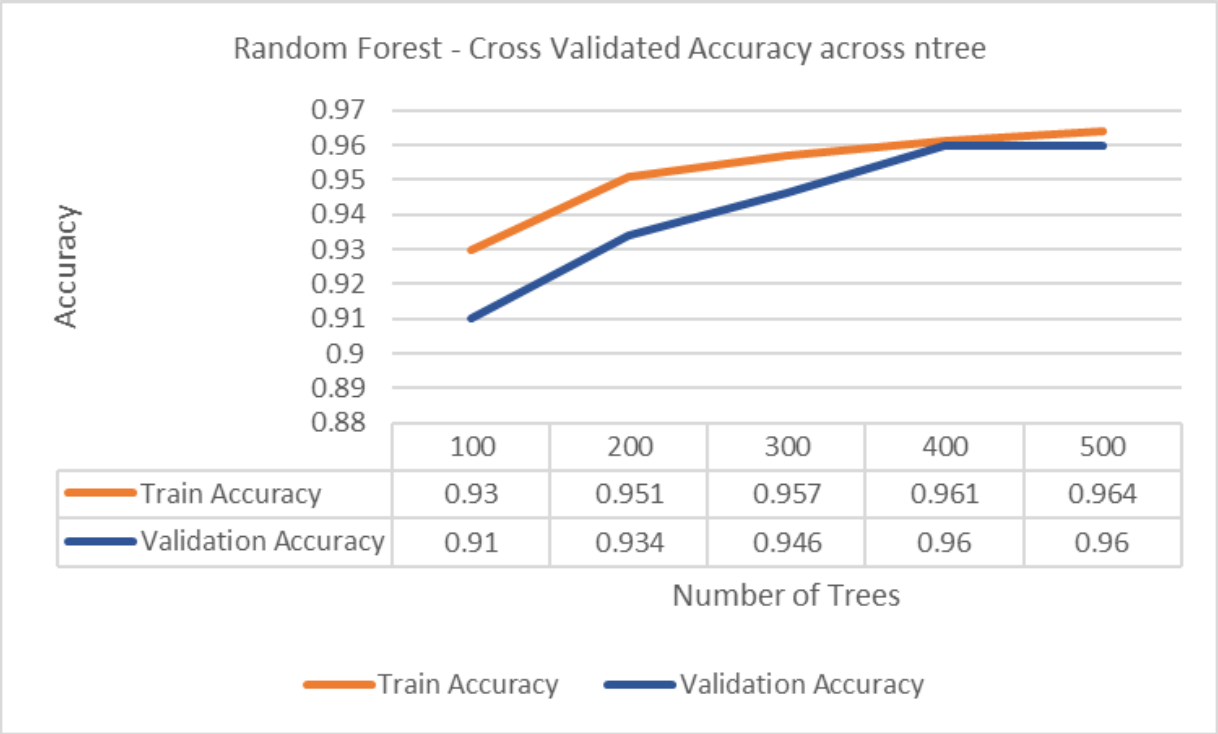
D. Random Forest:

Random Forest[3] is an ensemble learning method known for its robustness and accuracy, primarily used for classification and regression tasks. It operates by constructing multiple decision trees during training and outputs the mode of the classes (for classification) or mean prediction (for regression) of the individual trees. This method is effective due to its ability to handle large datasets with high dimensionality and provides inherent feature importance measures, while also being less prone to overfitting compared to individual decision trees.

Hyperparameter Tuning: We tune the number of trees (ntree) in the RF. We test out values between 100 to 500 at increments of 100.

Cross-validation: We perform 5 fold cross validation.

The training and validation accuracies for different values of ntree are as follows:

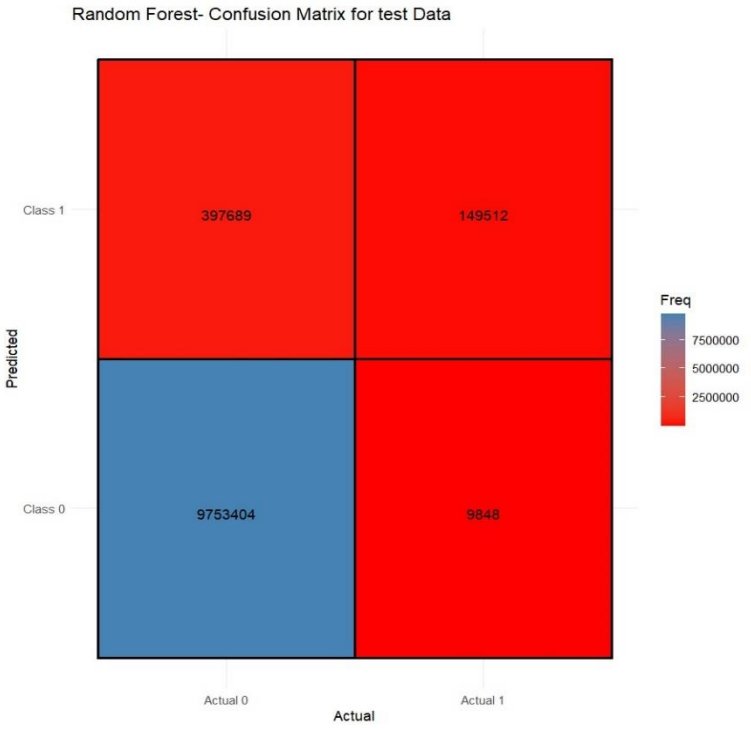


	100	200	300	400	500
Train Accuracy	0.93	0.951	0.957	0.961	0.964
Validation Accuracy	0.91	0.934	0.946	0.96	0.96

There's a significant increase in training and validation accuracies between ntree=100 and ntree=400 yet increasing beyond 400 does not significantly increase the training and validation accuracy but increases the computational complexity. Therefore, we choose 400 as our final value for the hyperparameter ntree.

Confusion Matrix on training data (RF):

Confusion Matrix for testing data (RF):



RF Evaluation Metrics		
	Accuracy	Recall
Training Set	96.14%	96.18
Testing Set	96.04%	96.08

Inferences

Upon comparing the accuracy and recall scores of K-Nearest Neighbors and Random Forest, we find that the Random Forest outperforms the KNN. The key contributing factors to the cancellation of the flights are the month of the year and certain carriers have a higher cancellation rate than others.

Managerial Discussion

Efficient operations are of utmost importance in the aviation industry. Predicting and preventing flight cancellations can greatly benefit airline managers. Our predictive model, powered by machine learning, offers valuable insights that can help enhance operational resilience. In this discussion, we will focus on the practical implications of our research and suggest ways for airline managers to effectively leverage the predictive model within their operations.

A. Proactive Operational Management

The predictive model developed in this research provides a powerful tool for airlines to proactively manage and mitigate flight cancellations. By identifying the key contributing factors, such as weather conditions, airport-specific challenges, and historical patterns, managers can implement targeted strategies to minimize disruptions. This proactive approach allows for the allocation of resources more efficiently, reducing the cascading effects of cancellations on subsequent flights.

B. Resource Optimization

Airlines can improve their resource allocation by analyzing the factors that cause flight cancellations. This analysis can help them to strategically assign staff, equipment, and maintenance resources based on predictive insights. By anticipating potential disruptions, airlines can create contingency plans that ensure that the necessary resources are readily available to address and resolve issues that could otherwise lead to cancellations.

C. Customer Service Enhancement

Flight cancellations can cause frustration and inconvenience for passengers. However, by utilizing the insights from our predictive model, airline managers can improve customer service by providing timely and accurate information to affected passengers. Proactive communication about potential disruptions will allow for better passenger management, alternative arrangements, and an overall improved customer experience, even in the face of unforeseen circumstances.

D. Operational Cost Reduction

Flight cancellations can lead to additional expenses for airlines, such as the cost of accommodating affected passengers and rearranging crew schedules. However, a predictive model can provide insight into potential cancellations, allowing airlines to optimize their operations and reduce costs. By

streamlining processes and allocating resources more efficiently, airlines can minimize the financial impact of disruptions and improve their profitability.

E. Continuous Improvement

The predictive model is not a fixed entity; instead, it can be constantly updated and enhanced using real-time data and changing operational patterns. Managers can create a feedback loop by integrating the insights and knowledge gained from previous cancellations into the model. This iterative approach guarantees that the predictive tool remains flexible and able to adapt to the ever-changing aviation industry landscape.

F. Regulatory Compliance

Airlines must work within a regulatory framework that stipulates penalties for flight cancellations. By utilizing our predictive model, managers can align their operations with regulatory requirements more effectively. This can help airlines minimize cancellations, which can lead to improved customer satisfaction and avoidance of financial penalties associated with non-compliance.

References

1. YUANYU 'WENDY' MU, “Airline Delay and Cancellation Data”, Kaggle.com, https://www.kaggle.com/datasets/yuanyuwendymu/airline-delay-and-cancellation-data-2009-2018/
2. Zhang Z. Introduction to machine learning: k-nearest neighbors. Ann Transl Med. 2016 Jun;4(11):218. doi: 10.21037/atm.2016.03.37. PMID: 27386492; PMCID: PMC4916348.
3. Breiman, L. Random Forests. *Machine Learning* **45**, 5–32 (2001). https://doi.org/10.1023/A:1010933404324