Abstract

The aim of this report is to apply a multitude of machine learning pipelines to an airport dataset. The features of this dataset, explore the various circumstances that a passenger may experience in an airport (parking, check-in, lounge etc.) and extracts a rating out of 5 for each of the experiences. The target variable is a rating out of 5, that a passenger has experienced. In this report, the Linear Regression, Decision Tree, LPBoost and Support Vector Machine pipelines will be analysed.

Contents

Abstract 1
Introduction 2
Domain Analysis 2
Data Cleaning, Pre-processing, and Feature Engineering 3
Dimension Reduction and Preventing Bias 4
  Dimension Reduction 4
  Correlations 6
Model Evaluation 6
  R Squared 6
  Mean Squared Error 6
  Root Mean Squared Error 6
The Machine Learning Algorithms 7
  Decision Trees 7
    Results 7
  Linear Regression 7
    Results 7
  Support Vector Machine 8
    Results 8
Cross Validation 8
Learning Curves 9
  Linear Regression 9
  Decision Tree 10
  LP Boost 11
  Support Vector Machine 11
  Generating the Learning Curve Graph using Matlab 12
Discussion 12
Conclusion 12

Introduction

Since the invention of the airplane, by the Wright Brothers in 1903, airports have become an integral part of our everyday journeys in airplanes. In the beginning, airports acted as simple portals for passengers to make their transition from land-based travel to air travel and only hosted minimalistic facilities to service planes and passengers. With the changing demands in aviation, increases in the number of passengers who fly, the changing security landscape around aviation and advances in airport architecture, airports have come to become the mini cities we know today.

Domain Analysis

The earliest airports started out as grass fields, where airplanes can land in any direction, if the wind is favourable. Hamburg Airport, which opened in January 1911 was the oldest international airport in the world. The airport supported both airship and aircraft operations and opened to international flights in 1955. Modern airports like London Heathrow and Dubai International have multiple, multi-level terminal buildings, often connected by rapid transit systems and custom developed people movers. Being the busiest airports in the world, both these airports can see aircraft take off within seconds of each other. The larger and more modern an airport is, it may have more passenger amenities, retail space, but at the same time can be very busy. This can have a two-way impact on passenger satisfaction. [1]

Between the years of 1961 and 1972, aircraft hijacking became even more common. Some bad actors used hijacking as a publicity tool, such as during the Cuban Revolution, hijackers wanted airlines in the US to fly to Cuba, thinking that they will be welcomed as revolutionary heroes by the new Cuban government.

The US government's answer to this, since January 1973, was the use of X Ray Machines and Metal detectors in airports. This technology has been historically used in prisons and military facilities across the US before, with incredible effectiveness. Both these systems were specifically used for detecting guns and other types of weapons.

Airport security can be a stressful process for most passengers, as they must stand in long queues, remove prohibited items from their luggage, this can negatively impact passenger satisfaction. [2]

Since the September 11th attacks in 2001, the US has signed the Aviation and Transportation Security Act (ATSA), which also established the Transport Security Administration (TSA), which was responsible for overall security across airports in the United States. The objective behind this was to instil confidence in air travel, which diminished right after the September 11th attacks. The drawback of this hypothesis was that passengers will have to sacrifice their personal freedoms and go through extra procedures at the airport before they get to board an aircraft. This had the potential of putting off passengers from flying all together.

As a result of the new security changes, passengers were instructed to arrive 2 hours before a domestic flight, where some passengers were asked to go through additional security checks. In December 2001, after a Shoe Bomb terror threat, passengers were also ordered to remove their shoes, when passing through the security check points. In some airports, the shoes themselves are sent through the scanner.

The intensity of TSA security checks can vary from passenger to passenger; hence some may have a more stressful process, whilst others may have a less stressful process. Being a US citizen also can make a big difference, as passengers can obtain preclearance and not have to go through security. Therefor the overall impact on passenger satisfaction may be varied. [3]

Once a passenger clears airport security and passport control, they are met with a maze of retail and duty-free outlets. First implemented at Shannon Airport in 1948, airport retail accounts for up to 40% of airport revenues; apart from aeronautic revenues (these are charges paid by airlines, such as airport fees, fuel, catering etc.).

By increasing airport retail revenue, airports can subsidize aeronautic costs, further making these airports attractive to airlines and passengers.

Many airport retail outlets follow an incomprehensible path, getting passengers to take a non-straight forward path to their gate. This gets passengers to spend more time in the retail section and maximises the visible retail space (this is the part of a retail outlet that is visible to the passenger, without the passenger having to go into the outlet itself.). Both these techniques can increase the amount that passengers spend in these outlets, thereby increasing revenues.

Airport retail can be either a hit or a miss with passengers, due to the variations in circumstances around them. Passengers can be more satisfied if they can bargain a deal with certain items, that cost more outside of an airport. But in cases where some passengers have consumed excessive duty-free alcohol products, they can ruin the overall experience for other passengers. [4]

Data Cleaning, Pre-processing, and Feature Engineering

There were 34 variables with missing values, where the following fields had the greatest number of missing values:

- Customs Inspection (5%)
- Speed of Baggage Delivery (5%)
- Overall Satisfaction (5%)
- Arrivals Passport and Visa Inspection (5%)

Since all the variable with missing data were Categorical Variables; where the values were always between 1 and 5, a **movmedian(n)** data filler was used. This takes a moving median from the previous and next ***n*** rows of the dataset. This was done with the assumption that the airport will have similar conditions at a certain window of time.

It was also observed that the **Departure Time** had rows with both the 12 hour and 24-hour time format, so this was converted such that the entire column had the 24-hour time format. The **hours** and **minutes** were then split into separate columns, this further allowed the column to become categorical. The understanding behind this is that passenger experience could correlate to certain parts of a day (for example the lowest score for security checks could be observed between midnight and 8am, due to potential short staff and tired passengers.).

The **Date Recorded** was also split into **day**, **month,** and **year**. This would help observe how passenger experience varied during the busier periods at the airport, such as the summer vacation and the Christmas break.

The **Quarter** column was dropped because this was covered in the **Date Recorded** column.

Dimension Reduction and Preventing Bias

Dimension Reduction

Principal Components 1 to 4 showed the greatest variance in the dataset. The following features contributed the most, in terms of variance, to the dataset:

ambienceofairport
arrivalspassportandvisainspection
availabilityofbaggagecarts
check_inwaittime
cleanlinessofairportterminal
comfortofwaiting_gateareas
courtesyofinspectionstaff
courtesyofofcheck_instaff
courtesyofsecuritystaff
customsinspection
daterecorded_day
daterecorded_month
daterecorded_year
departuretime_hour
easeofmakingconnections
efficiencyofcheck_instaff
feelingofsafetyandsecurity
groundtransportationto_fromairport
parkingfacilities
*parkingfacilities_valueformoney_*
speedofbaggagedelivery
thoroughnessofsecurityinspection
waittimeatpassportinspection
waittimeofsecurityinspection

The following bullet points explains the significance of these features:

- Cleanness of the airport terminal
    - As passengers spend long hours inside airport terminals that can get quite busy. Irregularly cleaned airports can have a negative impact on passenger experience. With

pandemics like COVID, it is more so important to keep airports clean, to reduce the spread of the virus, which can risk profits to the aviation industry.

- Ambiance of Airport
    - As air travel can be stressful, ambiance of the airport can have a significant impact on the passengers using it. A calming airport would have a better ambiance, than one that is noisy and unpleasant looking.
- Comfort of Waiting in Gate Area
    - Passengers who are travelling in economy, probably must spend time at the Gate Waiting Area before boarding their flights.
- Ease of Making Connections
    - As airports can get busy with many flights landing and taking off at any given point in time, there is a significant percentage of passengers that are connecting passengers (changing from one flight to another). As schedules can be tight, the ease of making connections is important to passengers for them to not miss their flights.
- Immigration/Emigration and Security Inspection
    - Since the dawn of terrorism or criminal hijacking of aircraft, passengers tend to feel safer when the security inspection is thorough, but this can also lead to bad experiences for some passengers; as security inspections can be more stressful.
- Date and Time Series Features
    - These are supposed to have the highest variance as it increases with time.

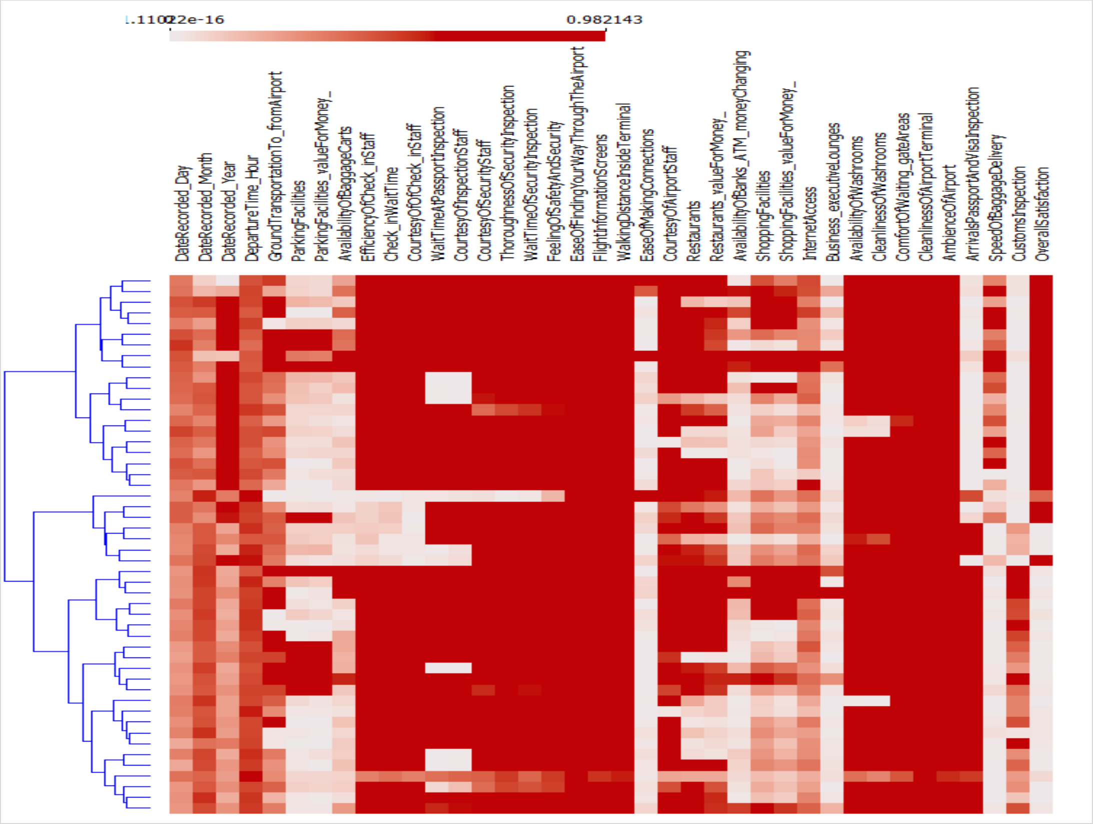

Figure 1 Heatmap for the PCA Analysis of the Airport Dataset

Correlations

Table 1 Features with high correlation to Overall Satisfaction

Feature	Correlation with Overall Satisfaction
DateRecorded_Year	0.473923101
SpeedOfBaggageDelivery	0.605141061
DateRecorded_Day	0.25122597
ArrivalsPassportAndVisaInspection	-0.895268385
CustomsInspection	-0.508436865

The positive correlation of 'DateRecorded_Year' and 'DateRecorded_Day' shows that the airport has been improving in terms of overall Satisfaction of Customers over time. The improvement of the speed of baggage delivery meant that, customers could exit the airport faster, spend more time on their holiday etc. thereby increasing overall satisfaction. Interestingly, the higher the rating for 'ArrivalsPassportAndVisaInspection' and 'CustomsInspection' the lower the Overall Satisfaction. This could be because Customs Investigations and Passport/Visa checks can be stressful on the passenger.

Model Evaluation

R Squared

R Squared is a measure of how a model's predictions closely reflect its inputs. This is usually a number between 0 to 1 and the closer to one, R Squared is, the better the model. [5]

$$R^2 = 1 - \frac{RSS}{TSS}$$

Mean Squared Error

The Mean Square Error is a measure of how closely a regression line follows its inputs. This is done by measuring the distance between these inputs and the regression line. [6]

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2$$

Root Mean Squared Error

Root Mean Squared Error is the standard deviation of the Mean Squared Errors. [7]

$$RMSE = \sqrt{\frac{\sum_{i=1}^{n}(X_i - \hat{X}_i)^2}{n}}$$

The Machine Learning Algorithms

Decision Trees

The decision tree is a type of supervised machine learning algorithm, where a model is developed in the form of an upside-down tree, where each split relates to a decision being taken and the end leave represent classes.

Results

Table 2 Decision Tree Results

Metric	Value
RMSE	0.057
MAE	0.012
R2	0.984

Linear Regression

Linear Regression is a type of supervised machine learning algorithm, where a linear model is developed based on the target and one or more features.

Results

Table 3 Linear Regression Results

Metric	Value
RMSE	0.121
MAE	0.090
R2	0.925

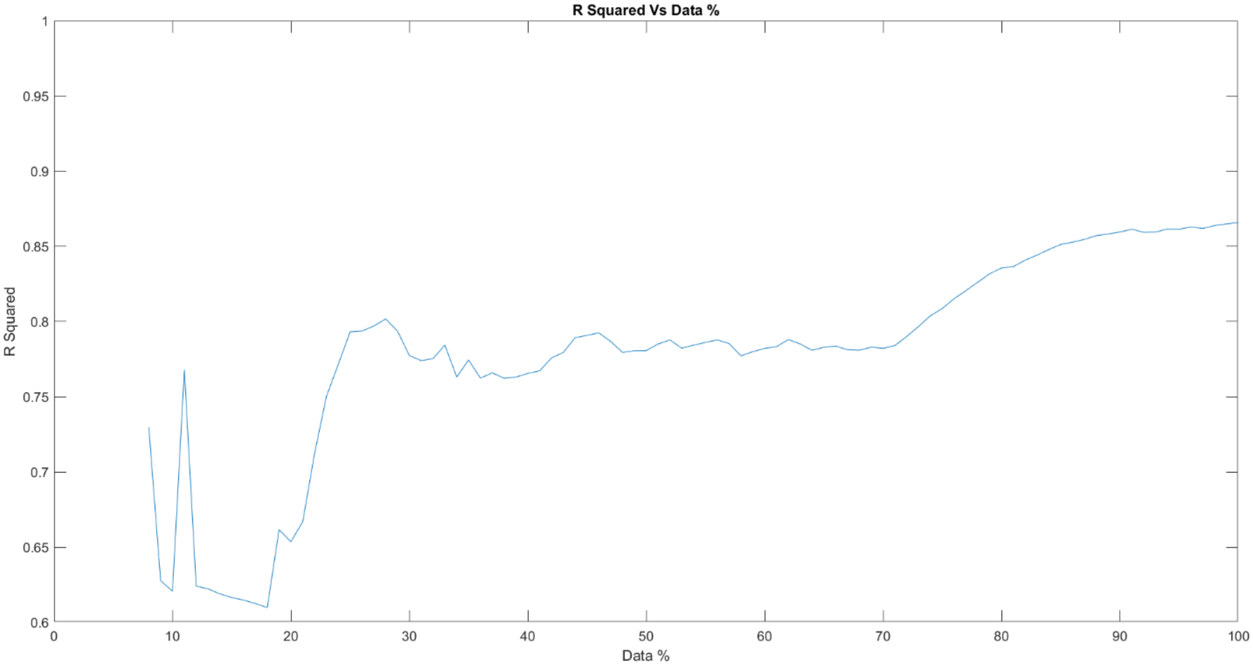

Figure 2 Graph of R Squared vs Percentage of data

Support Vector Machine

A Support Vector Machine (SVM) is a type of supervised machine learning algorithm, where hyper planes are defined in an **N** dimensional space, where **N** is the number of features in the dataset. The most optimal hyperplane can be defined by finding the maximum margins between two or more classes, as this helps the classification of future datapoints with more confidence. Support Vectors (SV) are data points that are closes to the margins of the hyperplane that helps decide the shape of a hyperplane. [8]

Results

Table 4 Support Vector Machine Results

Metric	Value
RMSE	0.145
MAE	0.116
R2	0.892

Cross Validation

Cross Validation is a methodology used in the machine learning pipeline, where a model can be validated for its efficacy. This used to be done by the traditional hold-out cross validation method, where a certain proportion is ear-marked for training and the rest for testing and the objective is to maximise both training and testing. The main problem with this method is that, for any higher percentage of training or testing data that is used, a sacrifice must be made for the vice versa and the training or testing component may not cover the entire dataset.

This is where k-folds cross validation comes in. K-folds works by splitting up the data set into ***k*** bins, where just like the earlier method, each bin will have a proportion of training data and testing data, but the sampling points of the train and test data will vary for each iteration. [9]

Learning Curves

Learning curves, depict the changes in training and validation errors, with the increase in number of data points. A wider gap between training and testing data, hints at the model being overly complex, whilst the lack of decent of the testing error, hints at inherent bias in the model. The following learning curves have been obtained from the various machine learning models that has been tested with this dataset.

Linear Regression

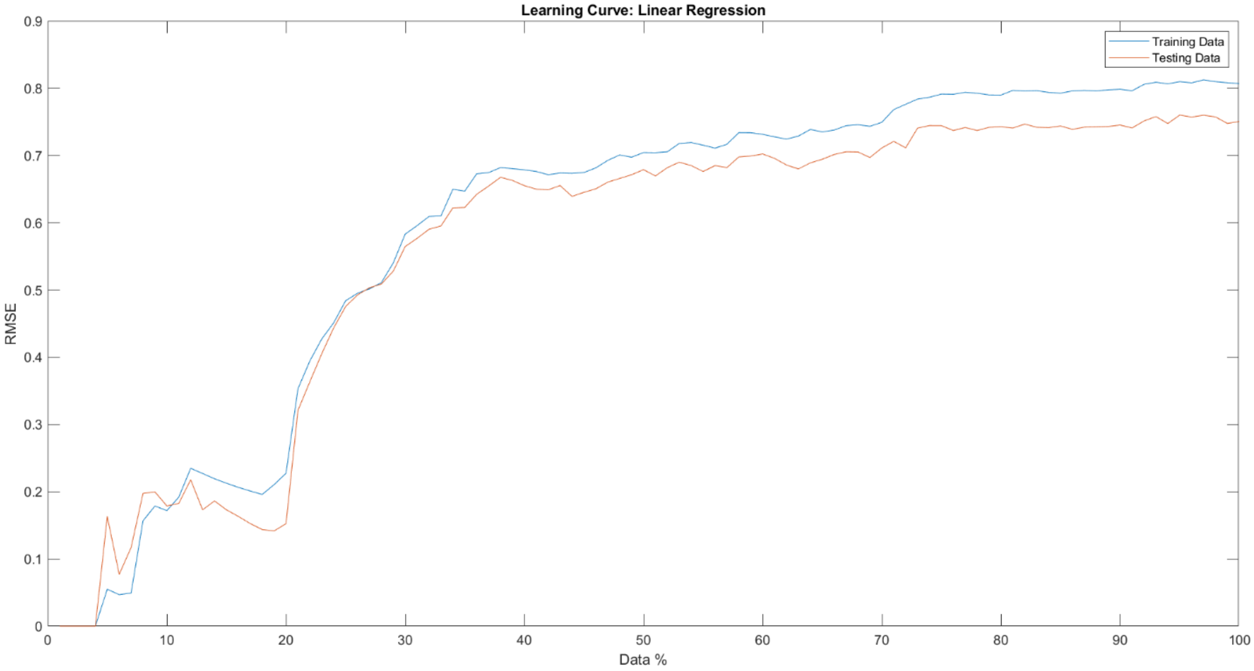

Figure 3 Learning Curve for Linear Regression

From the learning curve it can be observed that the model is not too complex or overfitting. This can be identified by the reasonable gap between training and testing data.

Decision Tree

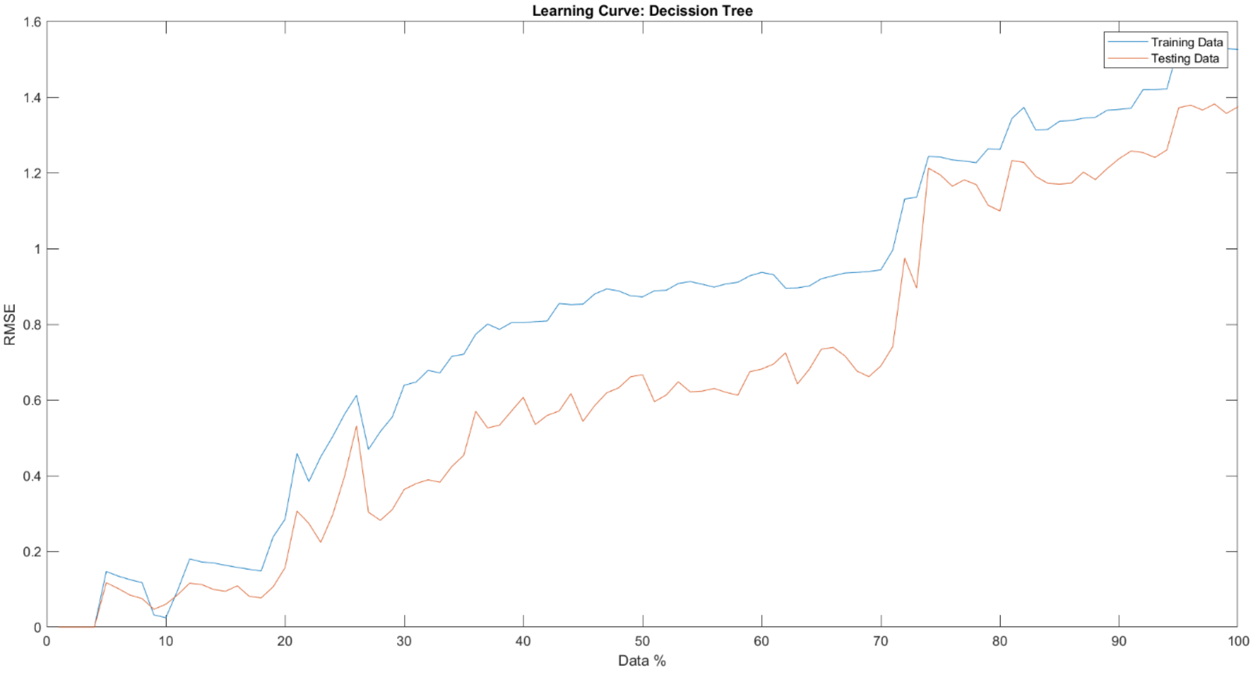

Figure 4 Learning Curve of Decision Tree, Max Splits =4

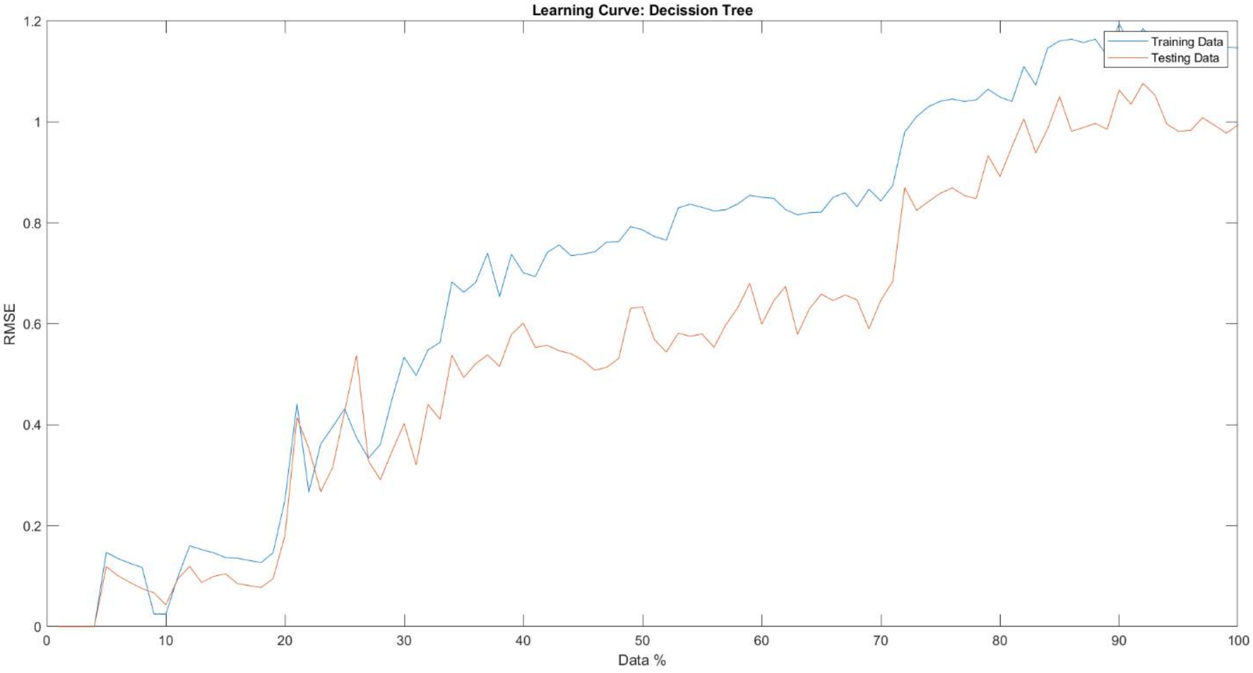

Figure 5 Learning Curve of a Decision Tree, Max Splits = 10

But varying the number of maximum splits in the decision tree, the complexity of the model can be varied. This results in the distance between training and testing data increasing proportionally with model complexity. As the amount of data being trained, increases, the overfitting decreases.

LP Boost

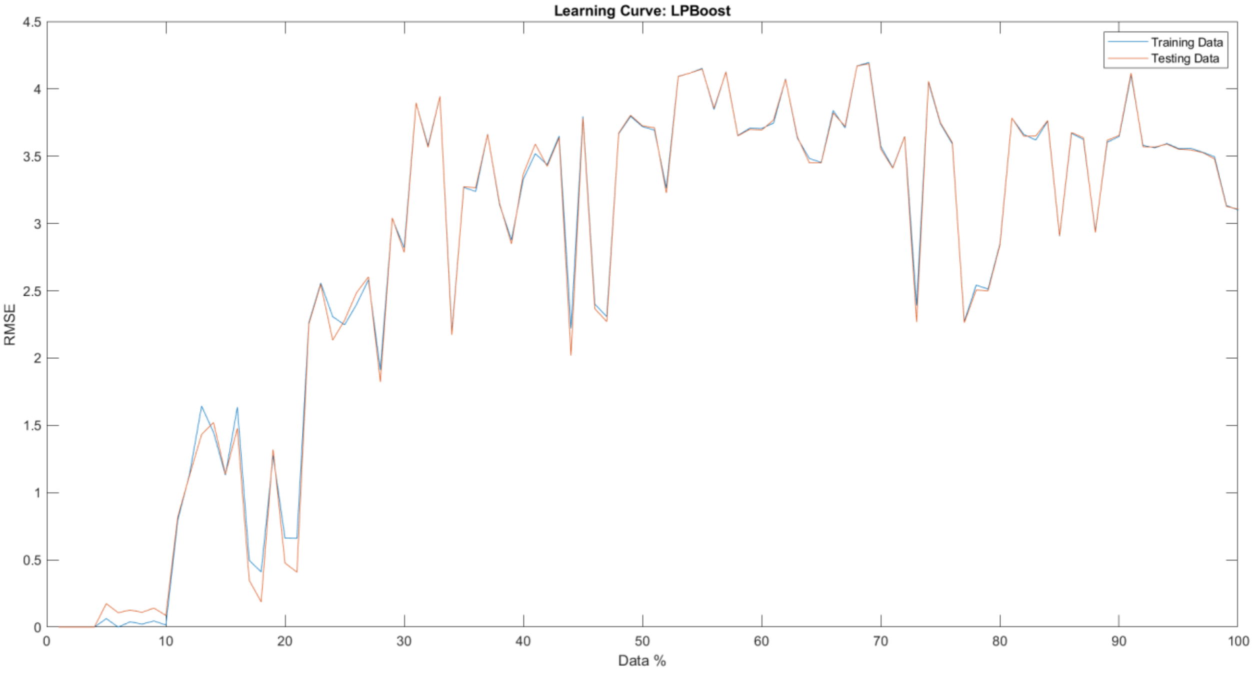

Figure 6 Learning Curve for LP Boost

In LPBoost, the model overfits and the gap between training data and testing data is negligible.

Support Vector Machine

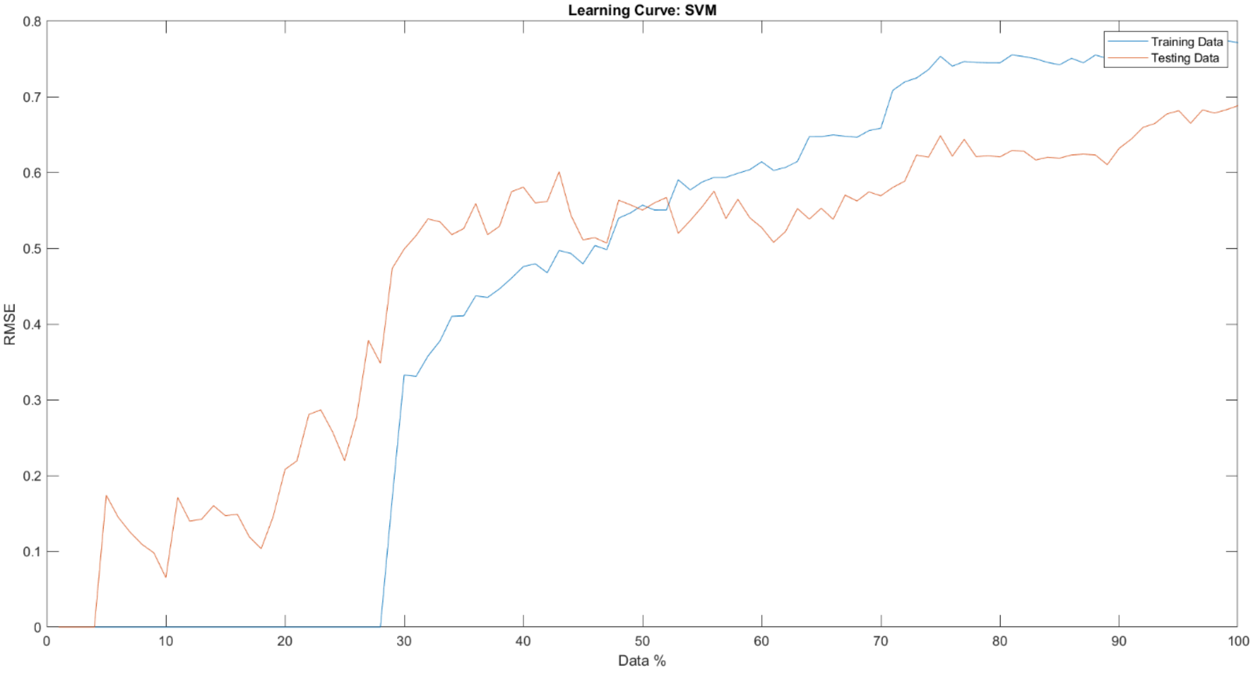

Figure 7 Learning Curve for the Support Vector Machine

As can be witnessed from the learning curve, the model is not overfitting as can be seen from the gap at the end. The model has reasonable complexity, as seen from a medium sized gap.

Generating the Learning Curve Graph using MATLAB

The learning curve in MATLAB is generated, by taking incremental percentages of data from the dataset and calculating their respective RMSE values. The graph of percentage of data vs RMSE values is then plotted.

Discussion

During the course of choosing, designing, and developing the MATLAB machine learning pipelines, it was deemed necessary to use Parallel Computing Techniques such as multicore/multithread operations to speed up the pipeline. This was because some pipelines were taking hours to complete. MATLAB allows for this, effortlessly, with the use of the Parallel Computing Toolbox and the Parallel For Loop (*par for*). Parfors allow for each of the iterations to parallelly run on a separate thread.

Conclusion

It can be determined that the decision tree is the best algorithm that can work with this dataset. This is because, it has the highest R value of all other algorithms, paired with a reasonable learning curve.

References

[1] J. Hayward, "The World’s Oldest Airports," Simple Flying, 12 May 2021. [Online]. Available: https://simpleflying.com/worlds-oldest-airports/. [Accessed 17 December 2021].

[2] A. Hay, "A brief history of airline security, hijackings and metal detectors," IBM, 24 April 2019. [Online]. Available: https://www.ibm.com/blogs/systems/a-brief-history-of-airline-security-hijackings-and-metal-detectors/. [Accessed 17 December 2021].

[3] G. Blalock, V. Kadiyali and D. H. Simon, "The Impact of Post-9/11 Airport Security Measures on the Demand for Air Travel," Jstor, [Online]. Available: https://www.jstor.org/stable/pdf/10.1086/519816.pdf. [Accessed 17 December 2021].

[4] K. Spinks, "Airport Retail: A critical revenue stream," International Airport Review, 21 July 2016. [Online]. Available: https://www.internationalairportreview.com/article/23906/airport-retail-critical-revenue-stream/. [Accessed 17 December 2021].

[5] J. Fernando, " R-Squared," 12 September 2021. [Online]. Available: https://www.investopedia.com/terms/r/r-squared.asp. [Accessed 12 November 2021].

[6] S. Glen, "Mean Squared Error: Definition and Example," [Online]. Available: https://www.statisticshowto.com/probability-and-statistics/statistics-definitions/mean-squared-error/. [Accessed 12 November 2021].

[7] S. Glen, "RMSE: Root Mean Square Error," [Online]. Available: https://www.statisticshowto.com/probability-and-statistics/regression-analysis/rmse-root-mean-square-error/. [Accessed 12 November 2021].

[8] R. Gandhi, "Support Vector Machine — Introduction to Machine Learning Algorithms," towards data science , 7 June 2018. [Online]. Available: https://towardsdatascience.com/support-vector-machine-introduction-to-machine-learning-algorithms-934a444fca47. [Accessed 17 December 2021].

[9] J. Brownlee, "A Gentle Introduction to k-fold Cross-Validation," Machine Learning Mastery, 3 August 2020. [Online]. Available: https://machinelearningmastery.com/k-fold-cross-validation/. [Accessed 17 December 2021].