Predicting US Flight Cancellations

By: Jason Kwan and Kyle Leung (Lecture 2 Group C)

Table of Contents

01

Introduction

Flight context and data set overview

02

Methodology

Data cleaning and modeling

03

Results & Discussion

Final constructed model analysis

04

Limitations & Conclusions

Setbacks, assumptions and final words

01

Introduction

Flight context and data
set overview

The Flight Cancellation Problem

Flight cancellations pose quite a costly problem (both in $ and time) with both passengers and airlines

Especially, with the COVID-19 pandemic, predicting flight cancellations would be an incredibly help in mitigating this annoyance.

Flights Data Set

69225

Observations

Each observation represents a domestic flight operated by large air carriers in the US

45

Variables

Detailed information recorded with each flight (ex: destination airport, day, and scheduled flight time)

02

Methodology

Data cleaning and modeling

The Process

Clean Data

Deal with NA’s and multicollinearity

01

Model Data

Create several models from our data for predictions

02

Analyze Models

Find misclassification rates from testing

03

Compare Models

Choose model with best MCR and simplicity

04

Data Cleaning: Dealing with NA's

First, we needed to see if there were any predictors with NA values, and if there were, what was the proportion of NA's that each predictor has.

Predictors with >80% NA values

- Air system delay
- Security delay
- Airline delay
- Late aircraft delay
- Weather delay
- Share White/Black/Native/Asian/Hispanic
- Median Income
- Poverty Rate
- Percent Completed HS

Predictors with <10% NA values

- Tail number
- Pass traffic
- Aircraft Movement

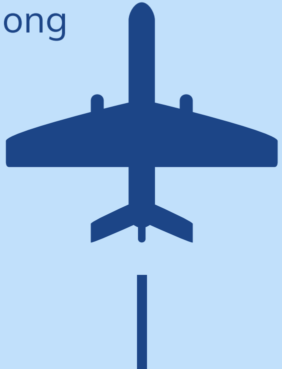
Concerning the predictors with >80% NA values, it would not be beneficial to try to impute these values since we are already missing so many already. So, we deleted these predictors entirely. We also removed the predictors with <10% NA values because they ended up being highly correlated with other predictors or had massive amounts of levels

Data Cleaning: Dealing with Multicollinearity

We found 3 main groupings of highly correlated variables. Afterwards, we chose and removed variables, accordingly.

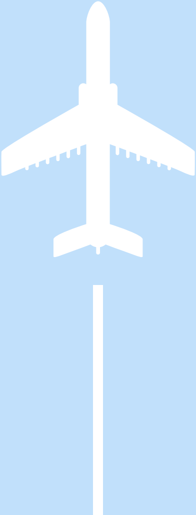
Origin

O.City
O.State
Origin_airport
Origin_city
Org_airport_lat
Org_airport_long



Time

SCHEDULED_DEPARTURE
SCHEDULED_TIME
SCHEDULED_ARRIVAL



Destination

Destination_airport
Destination_city
Destination_population
Dest_airport_lat
Dest_airport_long
Rank
Average.Passengers
Rank.Status

Cleaned Flights Data Set

69225

Observations

We were able to keep 100% of the original observations from the original data set

16

Variables

We were able to reduce the 45 original variables down to only 16, including cancellations

We were now able to move on to the modeling phase and try to create the best model at predicting flight cancellations

Data Modeling: Choosing our model

Logistic Regression

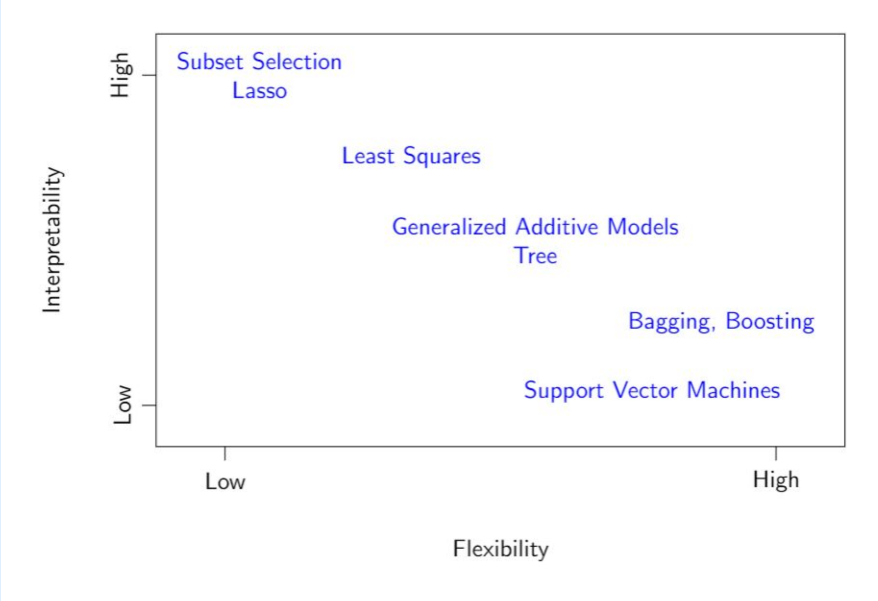
Pros:
Simple
Fast
Interpretable

Cons:
Inaccurate
Handles Poorly
Non-Flexible

Random Forests/Boosting

Pros:
Accurate
Handles Well
Flexible

Cons:
Complicated
Slow
Non-Interpretable



After splitting the data into training and testing sets (70/30), we decided to try both methods to see which one works better.

Logistic Regression Model

Method

We constructed a logistic regression model on predicting flight cancellations (YES for cancelled, NO for not cancelled) with the other 15 predictor variables on the training data:

Destination_airport, O.State, Passengers, Seats, Flights, Distance, Origin_Population, MONTH, DAY, DAY_OF_WEEK, AIRLINE, FLIGHT_NUMBER, SCHEDULED_DEPARTURE, SCHEDULED_ARRIVAL, and DIVERTED

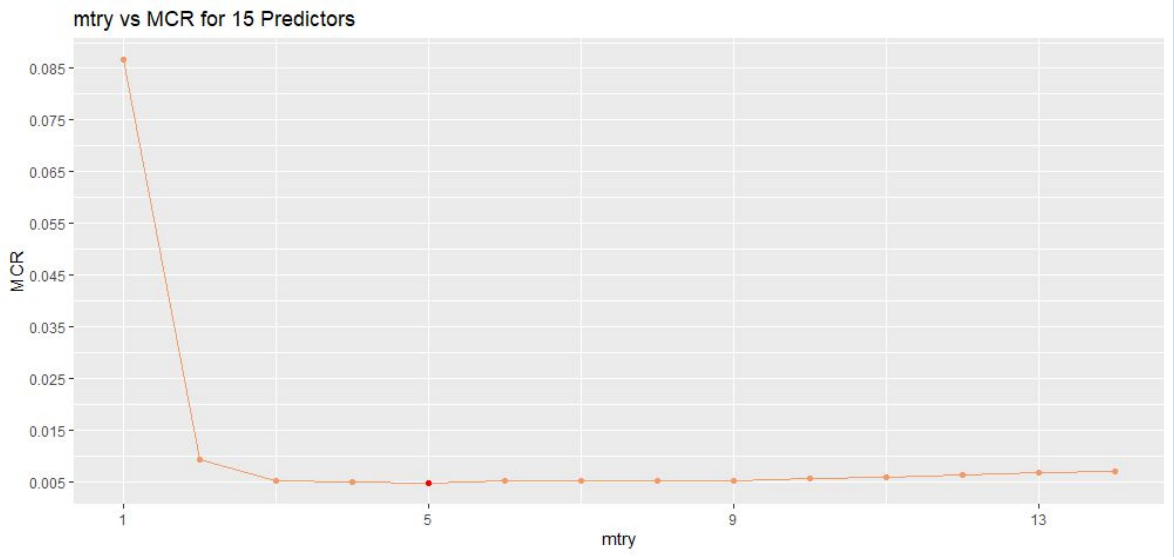
We then tried to predict cancellations on the testing data using the model.

Test Data Confusion Matrix (X: Predicted vs Y: Actual)

	NO	YES
NO	11859	3806
YES	1950	3152

We got a misclassification rate of 27.72%, which isn’t bad, but let’s see if forests can improve this!

Random Forest Model (Full)



Method

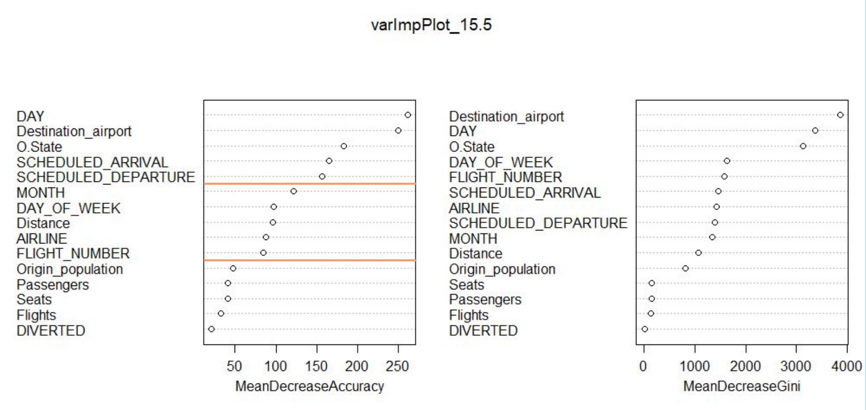
We constructed random forest models using all 15 predictors from training. However, we varied the mtry parameter from 1-14 to see which mtry works the best.

We then took all 14 models and tried to predict the testing data and compared the misclassification rates.

We can see that the MCR converges to about 0.58% with the minimum occurring at mtry = 5.

Random Forest Model (Full)

Variable Importance Plot
15 Predictors, mtry = 5



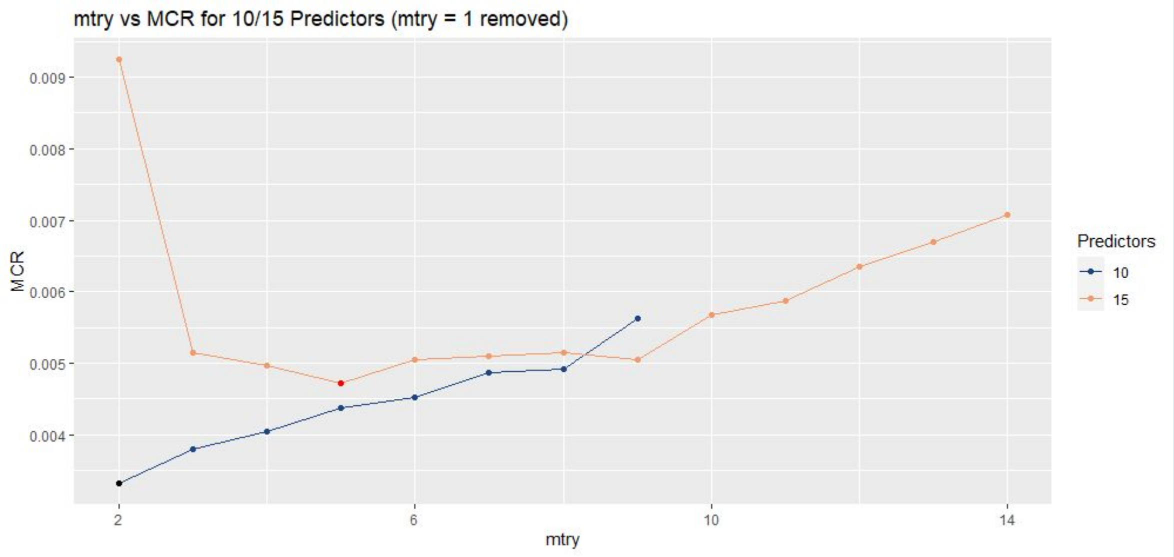
After looking through all of the variable importance plots, we noticed that the same predictors were grouped like above. Let's try simplifying our model by removing the predictor group with the lowest importance.

Test Data Confusion Matrix
(X: Predicted vs Y: Actual)

	NO	YES
NO	13735	24
YES	74	6934

We got a misclassification rate of 0.47%, which is a massive improvement from logistic, but let's see if we can get a similar rate but with fewer variables!

Random Forest Model (10P's)



Method

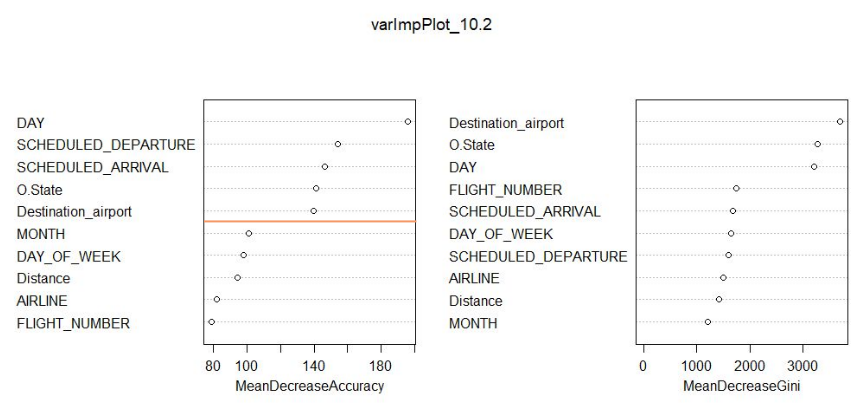
We removed the variable group with the lowest importance: Origin_population, Passengers, Seats, Flights, DIVERTED

Then, we repeated what we did before and varied mtry, now from 1-9 and compared the resulting misclassification rates.

Surprisingly, our new MCR's were actually better than the MCR's from our full model! (Minimum occurred at mtry = 2)

Random Forest Model (10P's)

Variable Importance Plot
10 Predictors, mtry = 2



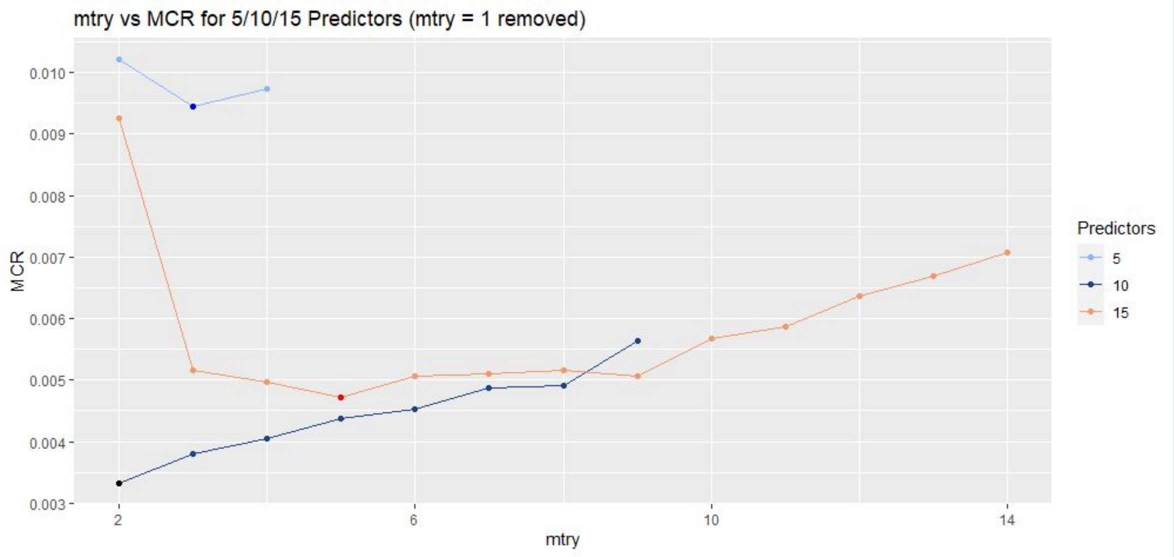
After looking through all of the variable importance plots, we noticed that the same predictors were grouped like last time. Let's try to simplify our model even more again.

Test Data Confusion Matrix
(X: Predicted vs Y: Actual)

	NO	YES
NO	13765	25
YES	44	6933

We got a misclassification rate of 0.33%, which is an improvement from last time but with even fewer variables! Let's see if we can do it again with even fewer variables.

Random Forest Model (5P's)



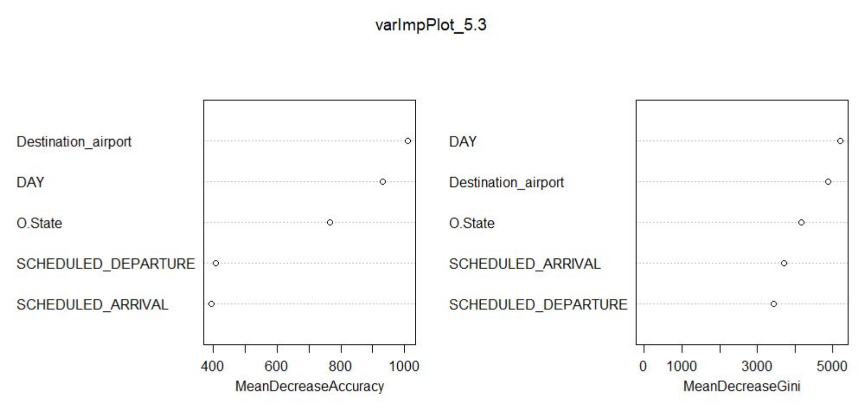
Method

We removed the variable group with the next lowest importance:
MONTH, DAY_OF_WEEK, Distance, AIRLINE, FLIGHT_NUMBER
Then, we repeated what we did before and varied mtry, now from 1-4 and compared the resulting misclassification rates.

Sadly, these new MCR's did worse and were actually far larger than even our full model MCR's. (Minimum occurred at mtry = 3)

Random Forest Model (5P's)

Variable Importance Plot
5 Predictors, mtry = 3



Even though, we can see an obvious separation here, we did not proceed with removing predictors further due to the large jump in MCR's.

Test Data Confusion Matrix
(X: Predicted vs Y: Actual)

	NO	YES
NO	13643	30
YES	166	6928

We got a misclassification rate of 0.94%, which was not an improvement from last time. However, we still have one more technique to try!

Boosting Model

Method

We constructed a boosting model with all 15 predictors. In terms of accuracy, random forests and boosting should theoretically be some of the best techniques at classification. So, we wanted to see if boosting does just as, or even better, than random forests.

After constructing the model from the training data, we then proceeded to, as always, try to predict cancellations on the testing data.

Test Data Confusion Matrix (X: Predicted vs Y: Actual)

	NO	YES
NO	13451	358
YES	48	6910

We got a misclassification rate of 1.96%, which is actually larger than all the MCR's from random forest. We then concluded our modeling phase and now we had to choose which model to use.

Final Model

Choosing the Model

When choosing the best model, we ideally wanted the model with the lowest MCR and is the most simple. However, realistically, this is not always possible and so we have to balance and weigh our options.

From seeing our model's MCRs and simplicity, we decided to choose our random forest model (mtry = 2) with 10 predictors.

It had the lowest MCR as well as it being "moderate" in terms of simplicity among our options.

Misclassification Rates

In %	Logis.	RF 15	RF 10	RF 5	Boost
MCR	27.72	0.47	**0.33**	0.94	1.96

Simplicity

	Logis.	RF 15	RF 10	RF 5	Boost
Rank	1	4	**3**	2	5

03

Results & Discussion

Final constructed model analysis

Analysis: Key Parts

Model

Random Forest (mtry: 2)

Observations

69225 Flights

Predictors

10 Flight Predictors

Simplicity

Moderate

MCR

0.25%

Rank

Top 15

Discussion: The Important Predictors

01

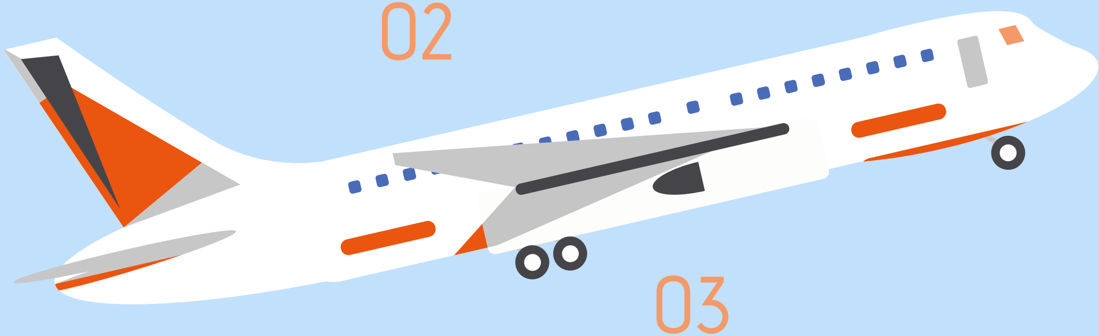
Location

Origin State, Destination Airport, and Distance

02

Date

Month, Day, and Day of the Week



03

Flight Characteristics

Airliner Name and Flight Number

04

Time

Scheduled Arrival and Departure

Discussion: The MOST Important Predictors

Origin State

Day

Destination Airport

From our model, we were able to find exactly what predictors were the most important in predicting flight cancellations. For the most part, if you know where the flight came from, where its going to, and what day it will happen, you should still be confident in being able to accurately predict whether the flight will be cancelled or not using our model, and be able to plan accordingly!

04

Limitations & Conclusions

Setbacks, assumptions and final words

Limitations

Data Cleaning

May have chosen weaker variables

01

02

Modeling

Was not able to use regular trees

Assumptions

Fortunately, there are no RF assumptions

03

Conclusion

We get very good prediction using our model, however, we get poor understandings of our causes. The typical variables weren’t included in our dataset (ex. Weather)

Further research would be necessary to delve deeper into this problem

But we believe passengers and airlines would still benefit from our model!

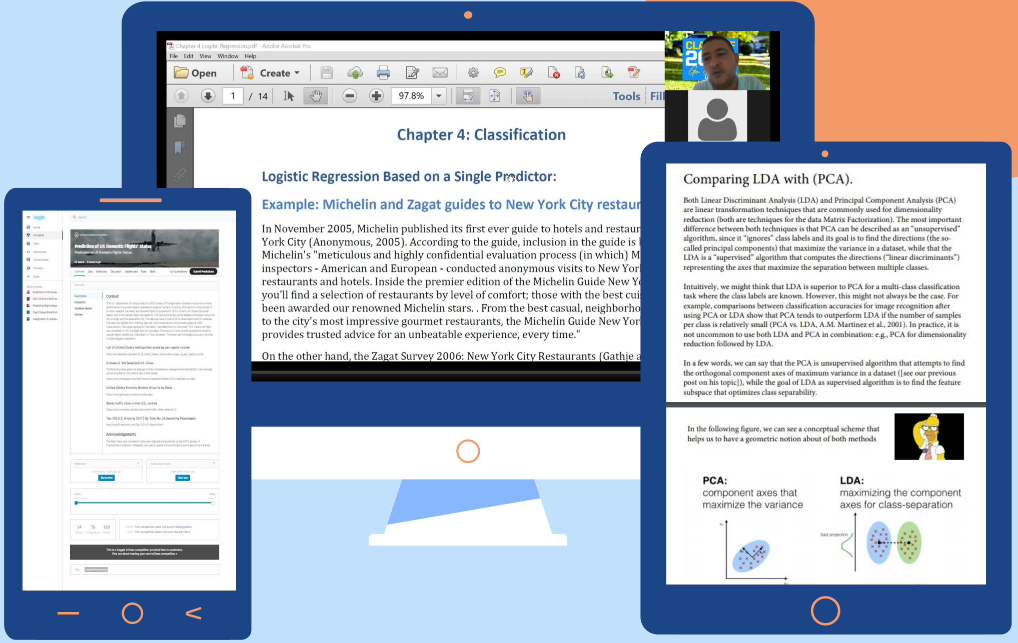
References

Special Thanks To:

Almohalwas and Wang's Stats 101C Lectures and Discussions

Towards Data Science Inc. https://towardsdatascience.com/

Biodiversity and Climate Change Virtual Laboratory https://bccvl.org.au/



Chapter 4 Logitic Regression.pdf - Adobe Acrobat Pro
File Edit View Window Help
Open
Create
1 / 14
97.8%
Tools
Fill

Chapter 4: Classification

Logistic Regression Based on a Single Predictor:

Example: Michelin and Zagat guides to New York City restaur

In November 2005, Michelin published its first ever guide to hotels and restaur
York City (Anonymous, 2005). According to the guide, inclusion in the guide is
Michelin's "meticulous and highly confidential evaluation process (in which) M
inspectors - American and European - conducted anonymous visits to New Yor
restaurants and hotels. Inside the premier edition of the Michelin Guide New Y
you'll find a selection of restaurants by level of comfort; those with the best cui
been awarded our renowned Michelin stars. . From the best casual, neighborho
to the city's most impressive gourmet restaurants, the Michelin Guide New Yor
provides trusted advice for an unbeatable experience, every time."

On the other hand, the Zagat Survey 2006: New York City Restaurants (Gathje a

Prediction of US Domestic Flights' Status

Predictions on US Domestic Flights' Status

34 teams · 6 hours to go

Overview

Context:

The U.S. Department of Transportation's (DOT) Bureau of Transportation Statistics tracks the on-time performance of domestic flights operated by large air carriers. Summary information on the number of on-time, delayed, canceled, and diverted flights is published in DOT's monthly Air Travel Consumer Report and in this dataset flight cancellations. The data set has also some detailed information about the city of origin and the destination city. The data set have a total of 2543 observations and 44 variables. The data was splited into a training data set (1653 observations) and a testing data set (890 observations). The target variable is "Cancelled". Cancelled has two outcomes "YES" means the flight was canceled or "NO" the flight was not canceled. The task is to come up with a predictive model to classify flight details into "Cancelled" or "Not Cancelled". The team with the largest accuracy wins the In-Class kaggle competition.

List of United States metropolitan areas by per capita income

https://en.wikipedia.org/wiki/List_of_United_States_metropolitan_areas_by_per_capita_income

Climate of 100 Selected U.S. Cities

The following table gives the average monthly temperature, average annual precipitation, and average annual snowfall for 100 cities in the United States.

https://www.infoplease.com/math-science/weather/climate-of-100-selected-us-cities

United States Airports: Browse Airports by State

https://www.globalair.com/airport/state.aspx

Worst traffic cities in the U.S., ranked

https://www.cbsnews.com/pictures/worst-traffic-cities-america10/

Top 100 U.S. Airports 2017 | By Total No. of Departing Passengers

http://www.fi-aeroweb.com/Top-100-US-Airports.html

Acknowledgements

The flight delay and cancelation data was collected and published by the DOT's Bureau of Transportation Statistics. Wikipedia was used to gather some information about airports and airlines.

There are no notebooks yet.

There are no topics yet.

Comparing LDA with (PCA).

Both Linear Discriminant Analysis (LDA) and Principal Component Analysis (PCA) are linear transformation techniques that are commonly used for dimensionality reduction (both are techniques for the data Matrix Factorization). The most important difference between both techniques is that PCA can be described as an "unsupervised" algorithm, since it "ignores" class labels and its goal is to find the directions (the so-called principal components) that maximize the variance in a dataset, while that the LDA is a "supervised" algorithm that computes the directions ("linear discriminants") representing the axes that maximize the separation between multiple classes.

Intuitively, we might think that LDA is superior to PCA for a multi-class classification task where the class labels are known. However, this might not always be the case. For example, comparisons between classification accuracies for image recognition after using PCA or LDA show that PCA tends to outperform LDA if the number of samples per class is relatively small (PCA vs. LDA, A.M. Martinez et al., 2001). In practice, it is not uncommon to use both LDA and PCA in combination: e.g., PCA for dimensionality reduction followed by LDA.

In a few words, we can say that the PCA is unsupervised algorithm that attempts to find the orthogonal component axes of maximum variance in a dataset ([see our previous post on his topic]), while the goal of LDA as supervised algorithm is to find the feature subspace that optimizes class separability.

In the following figure, we can see a conceptual scheme that helps us to have a geometric notion about of both methods

PCA: component axes that maximize the variance

LDA: maximizing the component axes for class-separation

Thank you for traveling with us!
And send us your questions!

THANKS

Merry Christmas and happy travels!

And always remember:

1. Don't drink and drive!
2. Don't shoot a mosquito with a missile!
3. And always know who the daddy is!

CREDITS: This presentation template was created by Slidesgo, including icons by Flaticon, and infographics & images by Freepik

Please keep this slide for attribution