Tyler Graham

CS 5610

December 11, 2023

Project 5

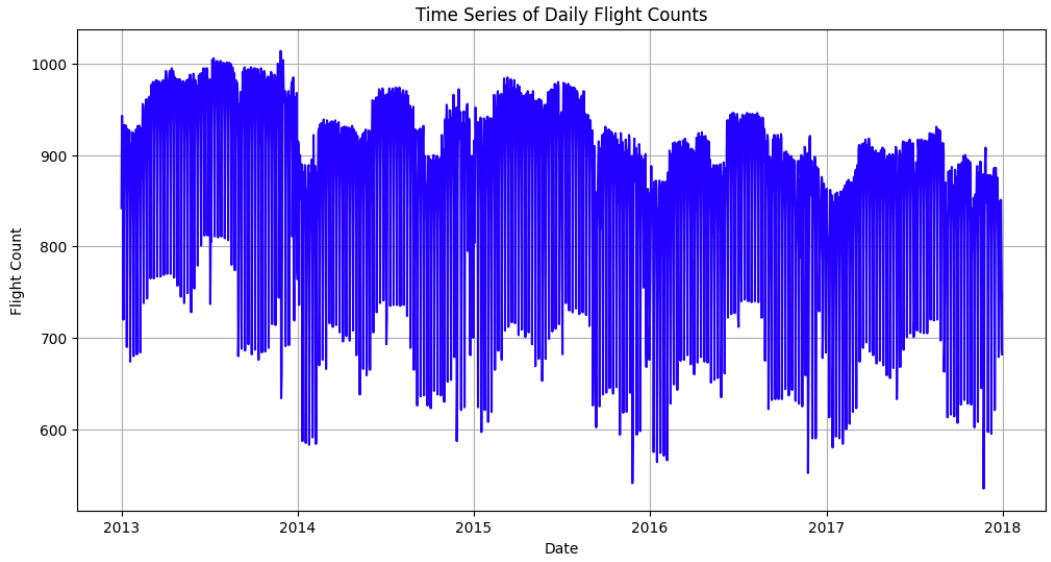
Project 5 involves a comprehensive examination of a series of datasets, specifically focusing on flight data spanning from 2013 to 2017 for New York City. These datasets, derived from annual records of flights, offer information on identifying trends, patterns, and operational characteristics in the aviation industry. The analysis aims to create a model to help predict the daily flights. This is key for airlines to predict demand and properly optimize both their airplanes, but support staff and airport staff. The initial data set includes the following columns.

Column Name	Type	Description
year	Integer	The year of the flight.
month	Integer	The month of the flight.
day	Integer	The day of the month of the flight.
dep_time	Integer	Actual departure time.
sched_dep_time	Integer	Scheduled departure time.
dep_delay	Integer	Departure delay in minutes.
arr_time	Integer	Actual arrival time.
sched_arr_time	Integer	Scheduled arrival time.
arr_delay	Integer	Arrival delay in minutes.
carrier	String	Airline carrier code.
flight	Integer	Flight number.
tailnum	String	Tail number of the airplane.
origin	String	Origin airport code.
dest	String	Destination airport code.
air_time	Integer	Duration of the flight in minutes.
distance	Integer	Distance of the flight in miles.
hour	Integer	Hour of the day when the flight was scheduled to depart.

Column Name	Type	Description
minute	Integer	Minute of the hour when the flight was scheduled to depart.
time_hour	String	Date and time in ISO 8601 format.

Each row in this dataset represents a single flight record, capturing all the details from the scheduled and actual departure times to the flight's origin and destination. The data includes dates from the 2013-2017 year period.

The first thing to do was to start to create a usable dataset to create some models. Due to the sheer size of the merged dataset of all of the flights per day from 2013-2017. A new data frame was created with each row being a new day and the flight count being the first row. This was plotted over data below. The key trends are every year during the summer months and spring break the flights increase and the first few months of the year are the down seasons.



In order to continue to develop a proper model, some additional features were created based of the original data.

Variable Name	Description
flight_count	The number of flights recorded on a given day.
avg_dep_delay	Average departure delay (in minutes) for flights each day.
avg_arr_delay	Average arrival delay (in minutes) for flights each day.
unique_carriers	The number of unique flight carriers on a given day.

Variable Name	Description
avg_air_time	Average air time (in minutes) for flights each day.
total_distance	Total distance covered by all flights on a given day.
flight_diversity	The variety of different flights (flight numbers) each day.
peak_departure_hour	The most common departure hour for flights on a given day.

A dataset with weather from the JFK airport was given for inclusion to in the data with the following parameters:

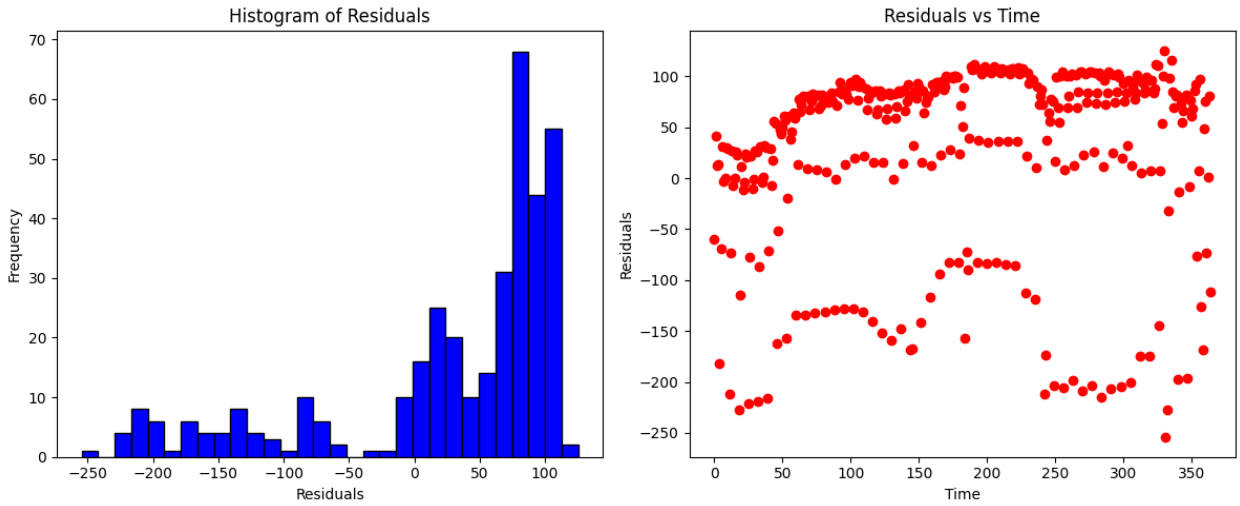
Variable Name	Description
PRCP	Amount of Rain
SNOW	Amount of Snow
TAVG	Average Temperature
TMAX	Maximum Temperature
AWND	Average Wind
TMIN	Minimum Temperature

In addition to the weather data, a categorical variable for if the date was a Holliday or not was used.

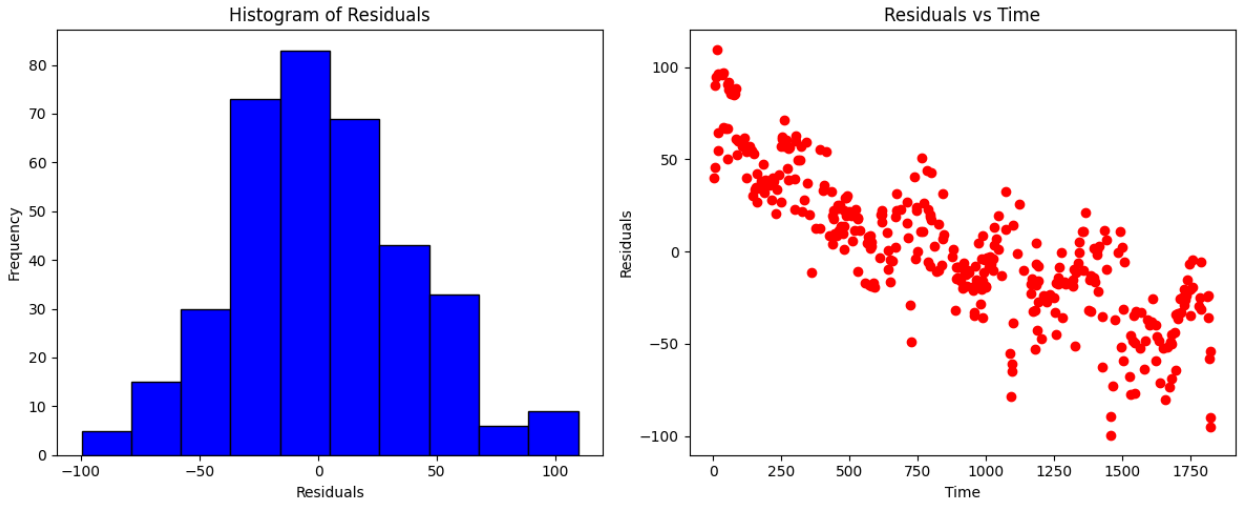
With the final data set three models were used for the initial development: Linear Regression, Ridge Regression and Random Forrest. The table below includes the following results of these models:

	Linear Regression	Ridge Regression	Random Forrest
MAPE	9.40%	3.50%	0.46%
RMSE	95.56	38.57	5.35

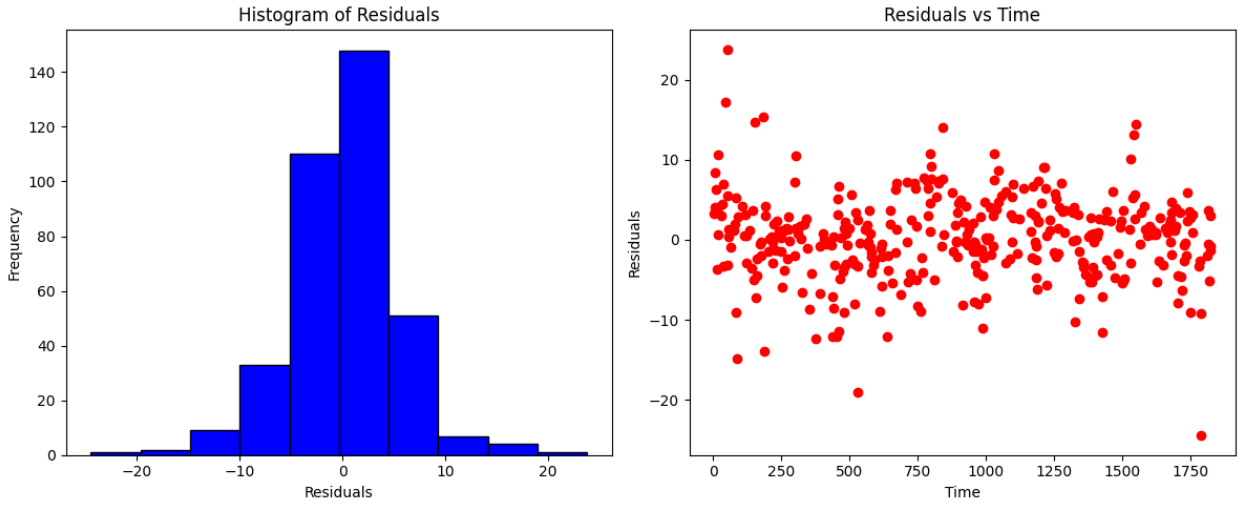
The alternative models each improved on each other as expected since they are more computational complex than a traditional linear regress model. These is also seen in the residual plots of each of these models below:



Linear Regression Residuals



Ridge Regression Residuals



Random Forest Residuals

In the analysis of residuals from the various models, it was observed that the residuals from the linear regression model are not normally distributed, suggesting that this model might be oversimplifying the data. This is evident from the large scatter in errors, indicating the model's inability to capture certain patterns or trends. On the other hand, the Ridge and RandomForest models exhibit normally distributed residuals, a sign that these models are better at capturing the underlying structure of the data without systematic biases. However, a notable pattern is observed in the Ridge model, where the residuals display a linear downward trend over time. This suggests that the Ridge model, despite its improved performance over simple linear regression, might still be missing some time-related aspect or trend in the data. In contrast, the RandomForest model's residuals show a more stable, horizontal pattern over time, indicating a consistent performance across different time periods. Its mean is very close to zero as well also suggesting that the performance is capturing the data correctly.

To better understand the models, the below chart contains the 10 biggest features for both models:

Rank	Ridge Feature	Coefficient	Random Forest Feature	Importance Score
1	total_distance	1.002896E-03	flight_diversity	0.992920
2	flight_diversity	1.292115E-06	total_distance	0.002589
3	avg_arr_delay	4.181870E-08	days_elapsed	0.002106
4	TAVG	1.266093E-08	peak_departure_hour	0.000306
5	avg_dep_delay	1.141878E-08	avg_dep_delay	0.000303
6	unique_carriers	1.006770E-08	avg_air_time	0.000292
7	TMAX	3.353133E-09	AWND	0.000259
8	AWND	1.917258E-09	TMIN	0.000256
9	IsHoliday_0	4.916971E-10	TMAX	0.000248
10	SNOW	7.604321E-11	avg_arr_delay	0.000247

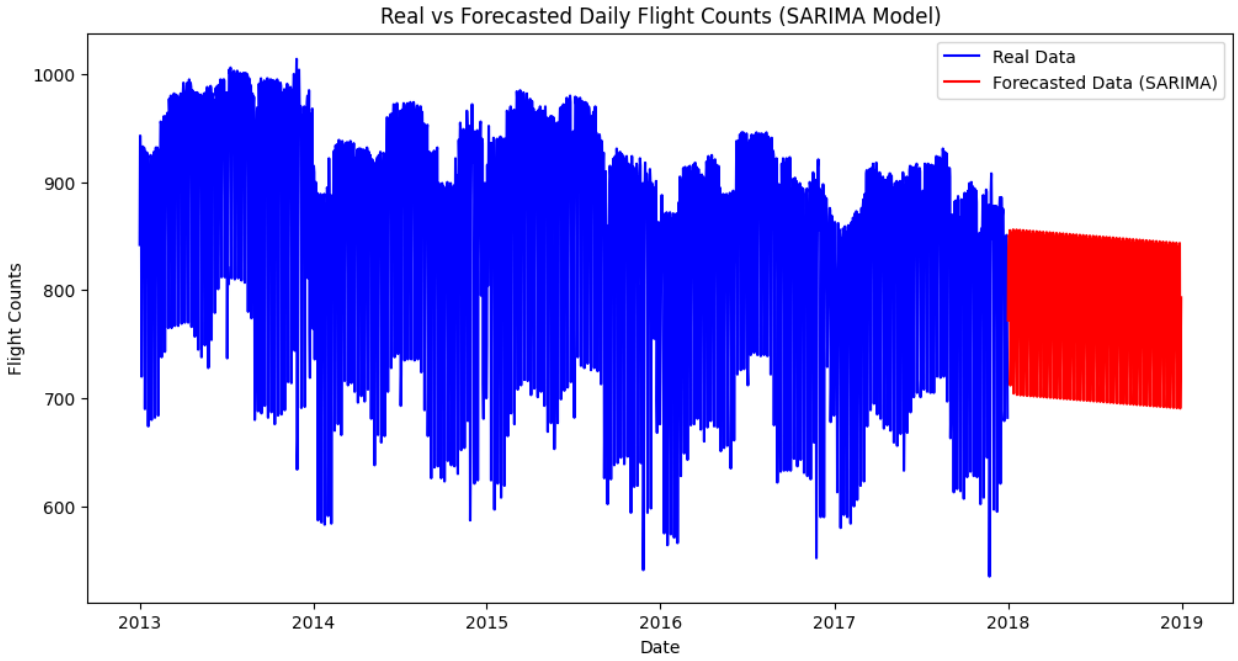
In Ridge Regression, a linear model, features are assigned coefficients indicating their linear relationship with the target variable. For example, total_distance has a significant positive coefficient, suggesting a direct and linear influence on flight count. Smaller coefficients, like for flight_diversity imply a more subtle effect. These coefficients denote the expected change in the target variable for a unit change in the feature, considering other features constant.

The Random Forest model uses feature importance scores to gauge a feature's usefulness in making accurate predictions. Unlike Ridge Regression's direct linear relationships, these scores reflect the overall utility of features in the model’s decision-

making process. For instance, flight_diversity scores high in importance, showing its crucial role in the model's predictions, despite not having a direct linear relationship with the target.

The differences in feature significance between the two models stem from their foundational principles. Ridge Regression's focus on linear relationships contrasts with Random Forest's handling of complex, non-linear interactions. A feature may be linearly influential but not as vital in non-linear contexts where Random Forest excels. This highlights the importance of choosing the right model based on data structure and prediction goals.

With the great performance of the models, the goal of forecasting the 2018 flight count was attempted. However, since we do not have the data for the weather and flight features a SARIMA time series model was chosen. The SARIMA model with parameters (1, 1, 1) for the non-seasonal component and (1, 1, 1, 7) for the seasonal component, it's clear that the results do not match what would be expected. This may be from not handling datasets with a lot of features or the parameters (p, d, q, P, D, Q) might not be optimal and need further tweaking. Re-evaluating the model parameters, considering different forms of seasonality, and possibly integrating external variables could provide more accurate and sensible forecasts.



Project 5 focused on analyzing flight data from New York City between 2013 and 2017 to predict daily flight numbers, a crucial factor for airline and airport staffing. The project involved processing detailed flight information into a daily summary, revealing seasonal trends with higher flights in summer and spring break, and fewer in early months. To enhance predictions, weather data and holiday indicators were included. The study employed Linear Regression, Ridge Regression, and Random Forest models, each showing progressively better performance, as evidenced in their residual

plots. While Linear Regression residuals suggested oversimplification, Ridge and Random Forest models displayed normally distributed residuals, indicating a better fit.

Comparing the top features in Ridge and Random Forest models revealed differences due to their distinct approaches: Ridge focused on linear relationships, and Random Forest on feature utility in complex, non-linear contexts. Finally, forecasting for 2018 using a SARIMA model was challenging due to missing weather and flight data and potentially suboptimal model parameters. This highlighted the need for further model tweaking and the complexities involved in predicting flight patterns in a environment like aviation.