A Work Project, presented as part of the requirements for the Award of a Master's degree in Business Analytics from the Nova School of Business and Economics.

THE IMPACT OF MACHINE LEARNING MODELS IN THE PREDICTION OF AIR CARGO

DIOGO ALEXANDRE MARUJO PINHEIRO

Work project carried out under the supervision of:

Pedro Brinca

Luis Rodrigues

17-12-2021

ABSTRACT

Forecasting is key in the aviation industry. Airlines use software applications to make cargo predictions. This thesis explores the advantages of training Machine Learning models using data from an airline to produce cargo forecasts and it evaluates the quality of these forecasts in the context of the business problem. The dataset available is curated and fed into different regression models, using Python scripts. Results are benchmarked with the current model used by the airline. Most of the models trained in this thesis serve the airline with better cargo predictions than the current model does, at one hour before the scheduled departure time.

Keywords (Machine Learning, Airlines, Cargo, Forecasts, Aviation, Data Science)

ACKNOWLEDGEMENTS

A special Thank You to Nhã, because you are always (ALWAYS) there, whether you're giving me advice or giving me emotional support. Thank you to my parents for suffering and celebrating with me every moment of my life, and for always being on my side. Thank you to my dog Woody, who will probably miss the afternoons spent sleeping beside me. Thank you to my friends Pedro, Rui, Nádia, Filipa, Osman, Sérgio e Inês. I also want to thank Professor Pedro Brinca and Professor Luis Rodrigues for giving me the opportunity to work in the industry that I love and for having the availability to be my advisors in this Master's thesis.

This work used infrastructure and resources funded by Fundação para a Ciência e a Tecnologia (UID/ECO/00124/2013, UID/ECO/00124/2019 and Social Sciences DataLab, Project 22209), POR Lisboa (LISBOA-01-0145-FEDER-007722 and Social Sciences DataLab, Project 22209) and POR Norte (Social Sciences DataLab, Project 22209).

LITERATURE REVIEW

THE AVIATION INDUSTRY

Aviation is a very dynamic industry due to the numerous internal and external factors that have a direct impact in its structure. These factors include the liberalization of markets (Shobande and Akinbomi 2020), the constant technological improvements, climate change and the ever-evolving competition forces (Wittmer et al. 2021). All these elements will have a direct impact in the business models that characterize the industry, which, in turn, have implications in the airlines' strategies, affecting operations. The sector has shown a steady and continuous growth during its more than one hundred years history, despite the setbacks that influence the demand for air travel (Liu et al. 2021).

AVIATION BUSINESS MODELS

Full-service network carriers (FSNC) traditionally evolved from the state-owned legacy carriers and are usually characterized by operating a hub-and-spoke network. This type of network, by contrast with the point-to-point direct itinerary, increases the complexity of flight scheduling (Morlotti et al. 2020), which, according to Aggarwal et al. (2020), is the most challenging optimization problem in the operation of airlines. This type of operation has a direct effect on the operating costs of the airline, which present a higher value when compared to airlines with different business models (Berritella et al. 2008). Other relevant factors that directly influence the operating costs include the provision of different travel classes and the transportation different types of traffic, such as passengers, cargo, and mail. Yet, within the FSNC business model, there are different ways in which the direct decision-making of management has an influence on the output produced by the company. One of those ways is to decide on the traffic mixture among passengers, cargo, and mail. This will influence the type of aircraft fleet that the airline operates, along with the role of the revenue management department. The latter's ultimate responsibility is to maximize revenue by matching supply with demand. Differences in traffic mixture can be found, by, for example, comparing the distribution of revenue tonne-kms carried between the two FNSCs Delta Air Lines and Korean Air. The distribution among passengers, mail and cargo is, respectively, 91%, 1% and 8% for the former, and 46%, 1% and 53% for the latter (Doganis 2019).

Low-cost carriers (LCCs) are the result of the liberalization of markets and deregulation. Changes in the regulatory landscape have huge impacts on increased competitive forces in the industry. Easier market entry by being able to fly between two points, the shift in airline ownership restrictions from conservative to liberal, and the relaxation of price fixing regulations have all contributed to greater competitiveness in the industry due to the emergence of new carriers. Allied to this, technological factors have made possible for LCCs to guarantee an advantage to their non-low-cost competitors. Younger fleets decrease variable operating costs for LCCs, as developments in both engine design and airframe composite have contributed to more fuel-efficient and lighter aircraft. The business model of the LCCs integrates the rationale of not offering connecting flights, even though some companies have started to explore the advantages in doing so (Maertens et al. 2016). The purpose of exclusively offering direct flights without the need to connect in a hub is to avoid complexity and drive costs downwards (Fageda et al. 2015). Doganis (2019) points to the fact that the absence of a hub operating connecting flights makes possible to LCCs to achieve quicker turnarounds whilst on the ground, maximizing aircraft utilization, and, consequently, reducing unit operating costs. This makes difficult for LCCs to handle cargo, as it would increase turnaround times. Both FSCs and LCCs rely on the transportation of passengers.

However, not all business models rely on the transportation of passengers as a source of revenue. In fact, there are carriers which exclusively transport cargo and mail, usually referred to as freight carriers (FC). These are still sub-divided into integrators (in case their business model also makes the provision of ground transportation of goods until they reach their final destination) and all-freight (if it only flies the goods between two airports) (Doganis 2019).

THE IMPACT OF COVID-19 IN AIR CARGO TRANSPORTATION

The aviation industry is characterized by a cyclical nature with regards to its financial performance, worldwide. The pattern is well-defined: five or six years of profits are followed by three or four years of losses. Additionally, during the years in which the industry has been profitable as whole, this profitability has only been marginal, not even covering the cost of the capital in many cases (Doganis 2019). There are many reasons that contribute to the cyclical financial shape of the industry, having external factors, such as epidemics, been playing a crucial role (Morrell 2013). This was the case in 2003, with the outbreak of SARS. According to IATA, 39 billion US dollars of revenue were lost solely by airlines operating out of the Asia-Pacific region (Forbes 2020), revealing the true devastating power that the spread of viruses have in the sector. This was no exception with the case of COVID-19.

COVID-19 is a virus that was first reported in 2019 and it was rapidly identified afterwards in more than 200 other countries (Lai et al. 2020). The negative effect that COVID-19 has had across multiple industries worldwide is undeniable. The tourism sector was one of the most affected sectors, with companies having to take dramatic measures to adapt to the circumstances: some of them with negative effects, but others with positive effects due to the surge of new business opportunities (Sánchez-Rivero et al. 2021). The negative impact that the pandemic has had in the tourism industry, in which the aviation industry is directly and largely involved in, arises from the intrinsic nature of the former. This nature is characterized by the mobility of individuals between different regions and the interaction among them (Yang et al. 2020). Scheiwiller (2021) states that the aviation industry has been severely impacted by the pandemic of COVID-19. Whilst some airlines were able to apply and be granted state loans, others were pressured to file for bankruptcy and close activity (Gottipati et al. 2020).

Therefore, airlines saw cargo transportation as a business opportunity, vital to soften the impact of the loss of passenger revenue. On one hand, and more importantly, airlines were able to use their aircraft to carry vital supplies such as medication and vaccines (IATA 2021), which has played a significant role for society in the fight against the virus. For carriers, it represented about a third of the total income, an increase of more than 10% when compared to levels before the COVID-19 pandemic. This is contrasted with cargo capacity having decreased more than 40% in 2020 in relation to the previous year (Maneenop and Kotcharin 2020). On the other hand, due to restrictions that forced entire populations to be on lockdown, online consumption was expected to rise, which would lead the demand for air cargo to increase (Özden and Celik 2021).

REVENUE MANAGEMENT AND THE NEED FOR FORECASTING

The aviation industry plays a big part in globalization. World trade and economic development is immensely facilitated by the provision of air cargo (Pearce 2019). For airlines, along with passengers, it is an important source of operational revenue. The main role of the revenue management department of an airline is to sell the right product, at the right price, at the right time, to the right customer (Yeoman and McMahon-Beattie 2017), and this is valid for both passenger and cargo traffic. However, a critical phase that precedes the task of Revenue

Management, is forecasting. Forecasting future demand is critical for an airline to define its future strategy (Ramadhani et al. 2020). However, this is not an easy task. The volatility combined with the unpredictability of patterns in demand, makes its forecasting a difficult mission (Vinod 2021). COVID-19 has increased the complexity to the key management's task of matching supply and demand, by bringing irregularity to both (Shaban et al. 2021). Rizzo et al. (2020) recognize the existence of differences between passenger and cargo revenue management, where the latter is characterized by volume constraints, the variability of both volume and weight and the absence of overbooking penalties. Nevertheless, the results that arise from the prediction of demand and the decision of supply outputted is vital for maximizing capacity utilization to maximize efficiency, guarantee higher levels of revenue and minimize costs (Wong et al 2020).

DATA SCIENCE

There seems to be a doubt in relation to the exact definition of what data science is and what the role of individuals whose activity is directly related to this field entails. Vaast and Pinsonneault (2021) defines this field of study as one in which specific skills are combined with the use of complex digital technologies such as programming languages and artificial intelligence as a resort to extract, from large datasets, knowledge that can be transformed into actions. It is the combination of different processes that include the collection, preparation, curation and manipulation of data and the use advanced data analysis techniques on that data, to be able to extract information and highlight patterns unknown until then (Oracle 2021). However, the analysis of data is not the goal itself. The aim of extracting relevant insights from the data is to then be able to use it as an important resource to inform actions. In order to make decisions, companies are revising their culture about data, from its collection to its use, due to the huge potential of the applications of Business Analytics (Bayrak 2021). From the broad range of what Data Science and Business Analytics involve, a key segment is the field of Machine Learning. Reports have demonstrated the unlimited possibilities of the application of Machine Learning (Tsai et al. 2021). Machine Learning is considered to be the field of study that combines the skills set from Mathematics and Statistics and Computer Science (Dangeti 2017). The ultimate purpose is to create a model, based on the dataset available, that recognizes patterns in the dataset to make predictions on new observations.

DATA SCIENCE IN AVIATION AND CARGO OPERATIONS

The aviation industry in no exception when it comes to the trend towards the use of Data Science as an important tool to aid management in making decisions. In fact, Chung et al. (2020) recognize the abrupt increase in academic research that combines the keywords "data science", "data analytics" and "aviation". Figure 1 shows the growth over the past recent years of published articles combining those three keywords.

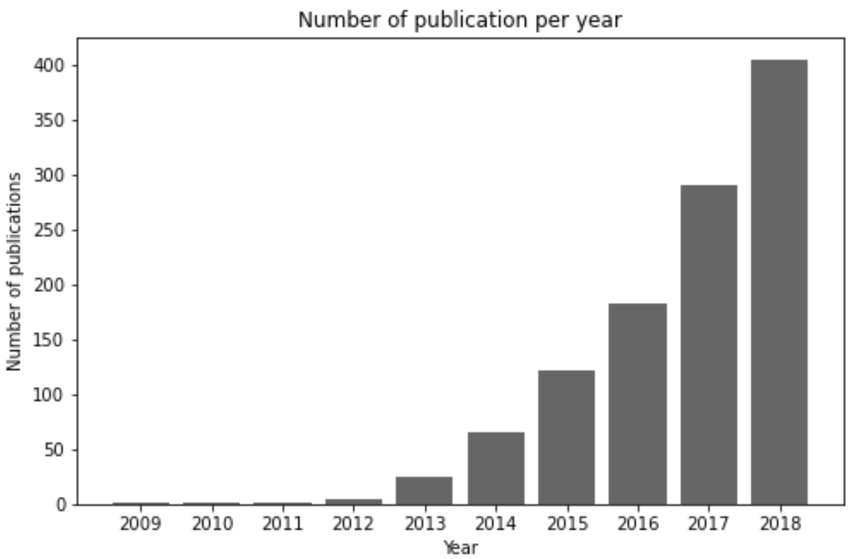


Figure 1 - The number of publications combining Data Science, Data Analytics and Aviation have been growing every year.

Chung et al. (2020) go further and subdivide the data science topic into different segments, from which Forecasting, Machine Learning and Air Logistics are an integrant part. Yet, these topics are very much interconnected. Forecasting is an aviation activity that can be done using both qualitative and/or quantitative methods. Quantitative methods use historical data to be able to predict into the future, typically using linear, exponential and regression models (Doganis 2019). For this reason, the use of data analytics skills is imperative. In the paper, Chung et al. (2020) refer to the Machine Learning topic as being the one that deals with very complex and dynamic problems. The study conducted by Niu et al. (2019) serves as an example of the implementation of Machine Learning models to make decisions in cargo airlines whose customer proposition rely on promised delivery time, suggesting that the investment and implementation of these models could not also benefit the airlines themselves but competitors also. Lastly, the subtopic of Air Logistics focus on its importance in modern operations management, having a critical role in quick response operations. Use cases in this topic reveal the use of models that develop metrics to evaluate the strength of networks (Zhou et al. 2019).

There is a need for airlines to increase efficiency and to reduce costs, and the use of Machine Learning has proved to be crucial in this task (Yaakoubi et al. 2020). Managing capacity, reduce operational costs and improve service quality are consequences of accurate forecasting, in which cargo volumes play an important role (Xu et al. 2019). This need is accentuated with the realization that, when it comes to cargo, an airline is competing with not only other airlines, but also with other means of transportation, such as trucks, trains, and ferries. Some of the relevant factors that contribute to the choice for cargo transport via air are the quickness, price, safety and security, quality of services, but, more importantly in the context of this thesis, the reliability and predictability (Morrell and Klein 2018). Janić (2019) highlights the importance for airlines to be resilient to the internal and external elements to be able to guarantee the quality of service to their cargo customers. Along with the general quality of the service, which has a positive effect in guaranteeing a long-term relationship with its clients, airlines should also be concerned with minimizing the costs associated with revenue management, i.e., balancing the risks of overbooking (Climato et al. 2020).

METHODOLOGY

RESEARCH QUESTION

The purpose of this thesis is to explore the extent to which the use of an airline's data to create Machine Learning models can yield cargo forecasts that can be translated into useful information to base decisions upon.

THE USE OF MACHINE LEARNING TECHNIQUES

Machine Learning is the process of programming scripts that can learn from an available dataset and with that develop a statistical model, as an attempt to solve problems. It is also commonly referred to as predictive analytics. This term is used because this branch of data science is trying to generate a generalized model, based on previous observations, to then apply to new, unobserved data, to be able to make predictions. However, not all research questions can make use of Machine Learning techniques to resolve a problem. There is a list of six questions that need to be verified, in the context of a business problem, to guarantee that Machine Learning is an applicable process in goal of seeking solutions.

The first question is whether or not there are clear, defined, and simple rules that can produce a forecast. Machine Learning is considered a valid approach in this case, as there are a considerable amount of variables with the potential to influence the answer, adding complexity to the model. The second question is related to the formulation of the problem. If the problem can be well defined, Machine Learning once again can be considered a relevant approach to solve it. In this case, the problem trying to be resolved is to make a prediction about an amount of cargo weight that will be loaded in an aircraft, one hour before departure, to use it as an aid for the decision-making process. Thirdly, the question that follows regards the presence of regular patterns in the data. For this question, an extensive exploratory data analysis is performed, with the data being curated accordingly when the need arises. The fourth question relates to the meaningfulness of the data. In this thesis, a description for each variable is presented and evaluated, which results in a key process of selecting the relevant features and of discarding observations that do not have a meaning in the problem scenario. The fifth question is how to define success. The predictions themselves cannot be evaluated if they cannot be compared to a benchmark value and/or or assessed in terms of the context in which they will be used in. The predictions will be contrasted with a pre-model generalized error to be able to evaluate the model and its contribution for the solution. The last question is related to the quantity of data. Machine Learning relies on relatively large datasets. The prepared dataset contains almost 100000 observations, each associated with a list of values for 12 attributes.

The category of the Machine Learning system applied in this thesis is “supervised learning”. This is because every observation has a set of values from the predictive attributes, which is associated with a value from an attribute that contains the desired value. The latter, also designated by ‘label’ variable, are what the model will predict based on new observations. Within the supervised learning system, the type of models utilized in this study have a regression nature. This is explained by the nature of the prediction, which, in this case, is a real number (the number of kilograms predicted of cargo to be loaded in an aircraft). Multiple

Machine Learning regression models have been utilized as an attempt to increase the chances of an improved result, which include Logistic Regression, Lasso Regression, Ridge Regression, Random Forest Regressor, AdaBoost Regressor and Neural Network. K-fold cross-validation is also performed during model training to better generalize the root mean squared error.

DATA COLLECTION

The dataset is the result of a concatenation of individual rows, each one representing one unique flight. The airline automatically gathers and stores the values of each variable in a database, combining data from different sources. These sources can either be input values manually inserted in a software, or a software that sends the data to the airline database. For the latter, before the software feeds the database with the data, pilots and other airline staff confirm the values. Nevertheless, there has been found some cases where the values of some variables are out of scope of what is feasible in the business context. Due to the inaccuracy of this data and the lack of opportunity to investigate the cause of those errors and to verify the correct values, these observations have been discarded from the dataset.

The dataset comprises both quantitative and qualitative data. The variables included in the dataset are variables that, in theory, could influence the weight of cargo to be transported in a flight. Quantitative data has been normalized before being used for training purposes. Quantitative data is related to the number of passengers on the flight, the filled capacity of the plane in percentage, and the cargo weight (in kilograms), that was both predicted and loaded into the aircraft. Regarding the qualitative data, i.e., the categorical variables, these assume different types of values from a limited range of possibilities. Qualitative features have been encoded and converted into binary variables.

Upon request, the airline has made this dataset available in the '.csv' format. A Non-Disclosure Agreement has been signed between three parties: the student, the University, and the airline.

SOFTWARE UTILIZED

The programming language utilized in this project is Python. The code run on Jupyter Notebooks, using Visual Studio Code. The data exploration and curation stages have been aided by using the 'numpy', 'pandas' and 'matplotlib' libraries, while the modelling phase has taken advantage of the code provided by the 'sklearn' and 'tensorflow' packages. The code utilized in both phases can be found in the Appendixes section.

LIMITATIONS

The limitations of the research focus on the time interval of the prediction. The airline is currently in possession of a software that makes a forecast of the available cargo weight that a specific flight will be able to carry. However, the forecasts are only available when consulted by an operator and are not recorded automatically in a database, making it only possible to produce forecasts of cargo weight one hour before departure time, in the context of this thesis.

Another issue is related to the lack of quality of the data. The data contains some errors that are easy to identify due to their unfeasibility in the real world. Even though it is claimed by the airline its human verification before being inserted in the database, the veracity of the values cannot be guaranteed just because the value is within an acceptable range of possible values. It is a fact that regardless of the models and hyperparameters used in the model construction, if the input data in invalid, the output of the models will always lack quality.

Lastly, the majority number of observations are related to a fleet that is relatively old in comparison with the usual life span of commercial aircraft. This means that even though the model results have the potential to be applied in business decisions in the short-term future, a fleet renewal could compromise the effectiveness of the Machine Learning model here presented, if the airliners acquired by the carrier are different from the ones being retired, as the model has been trained on aircraft types that will not be seen in new observations. The same happens if the airline adds new routes to their network.

RESULTS AND DISCUSSION

EXPLORATORY DATA ANALYSIS AND DATA CURATION

The first step in a data science project is to properly understand the dataset. Metadata is the designation given to the data that is used to describe and give information about other data. In this particular case, the metadata will describe the variables contained in the dataset. This is extremely valuable to have an overview of the data, to then plan the next steps in the data curation phase. Table 1 summarizes the metadata. The initial dataset contains a total of 130299 rows and 16 columns. Each observation contains information about a particular flight. For the purpose of this study, the target variable (label) is considered to be 'Cargo_actual', as this is the one variable trying to be predicted in future instances. It is important to have a holistic view of the quality of the data before this dataset is used to train a model. Having situational awareness about the business application of the model and its nature is key to ensure the dataset is reviewed and the reliability and quality of the data assessed. The process to curate the data will make sure that the dataset to be used to construct a machine learning model is exclusively composed of data with values in the right format, with no errors or discrepancies and that make business sense. Curating the dataset encompasses limiting the range of certain values, removing non-essential observations, and discarding outliers.

Table 1 - Metadata

NAME	DESCRIPTION	TYPE	MIN	MAX	MEAN	UNIQUE	MISSING
Date	The day of the flight (take-off time) at UTC time.	Numerical (Continuous)	——	——	——	——	0
Time	Scheduled departure time (UTC).	Numerical (Continuous)	——	——	——	——	0
Flight	Flight number.	Categorical	——	——	——	1402	0
Aircraft_type	Aircraft manufacturer and model.	Categorical	——	——	——	6	0
Origin	Origin three-letter IATA airport code.	Categorical	——	——	——	103	0
Destination	Destination three-letter IATA airport code.	Categorical	——	——	——	108	0
Adults	The number of passengers aged 12 years old or older.	Numerical (Integer)	-3	291	65.74	——	0
Children	The number of passengers aged between 2 and 12 years old.	Numerical (Integer)	0	92	3.01	——	0
Infants	The number of passengers aged up to 2 years old.	Numerical (Integer)	0	13	0.61	——	0
Load_factor	The percentage of sold seats in relation to the aircraft capacity.	Numerical (Float)	0	155.56	69.5	——	0
Pilot_ID	Number of identification of pilots.	Categorical	——	——	——	106	0
Fish_IO	It represents the absence (0) or existence (1) of fish in cargo.	Categorical	——	——	——	2	0
Animals_IO	It represents the absence (0) or existence (1) of animals in cargo.	Categorical	——	——	——	2	0
METAR_origin	The METAR weather conditions at the origin airport at scheduled departure time.	Categorical	——	——	——	5668	123376
Cargo_prediction	The prediction of cargo weight to be loaded (1h before departure time).	Numerical (Float)	0.0	10913.0	235.07	——	0
Cargo_actual	The actual cargo weight loaded.	Numerical (Float)	0.0	10969.0	298.09	——	0

By reading the information in table 1, one of the main inconsistencies found in the dataset is the number of missing values that the attribute 'METAR_origin' contains. From all the 16 features, this is the only that contains missing values, representing an extremely large number when compared with the total size of the dataset, accounting for almost 95% of the observations. This variable contains information about the weather conditions at the airport of origin, including wind direction, speed, and atmospheric pressure at mean sea level in the vicinity of the airport. Weather conditions are known to have significant implications in aircraft performance, and therefore, operations. This is manifested through cancellations, routing choice and aircraft loading decisions that affect the weight and the fuel consumption of an aircraft. Theoretically, this variable could have a relevant impact in the prediction of cargo to be loaded in a particular flight. However, due to the large number of missing values, no value-inputting strategy was implemented to replace the missing values. In addition to this, due to the inexistence of an official external source that could be used to retrieve the data, where its reliability could be verified, the decision made was to discard the variable 'METAR_origin' completely. Flights with duplicate entries have also been discarded.

The following process is the limitation of the range of some of the features. 'Load_factor' is the first attribute that needs modification. As described in table 1, it is an indicator of how full the aircraft cabin was after at take-off time. Its values can only range in a continuous scale from 0 to 100, as this is represented in percentage (passenger-wise, a flight can range from being empty to being full). The observations where a value greater than 100 was found were discarded. Incongruences have also been found in the category 'Adults', where the minimum value found was -3. The final dataset does not contain any observation where this variable presents negative values. Other variables presenting incongruences on its values are 'Origin' and 'Destination'. These present airports that do not exist, represented by the three-letter code 'XXX'. These observations have been discarded. Also, flights which share the same origin and destination airports have also not been included in the prepared dataset, as these represent flights forced to return to the origin airport (disrupted flights). Similarly, the feature 'Aircraft_type' presented values that require action on them. The fleet of an airline is not something static. In fact, throughout time airlines remove aircraft from their fleet either to retire them, to return them to the lessor or because the aircraft has been sold to another company. In the other hand, airlines make investments to purchase or to lease aircraft to replace aircraft being retired from the fleet, or as an investment to grow its current network and/or increase the frequency of flights on its schedule. However, even though using observations of flights operated by a certain type of aircraft in the past can be used in advantage to build a predictive model, as some airliners have been removed from the current fleet and the airline has no intentions of replacing them by the same equipment type in the short-term future, these will never be part of the new flight observations from which the model will make a prediction from. For this reason, the author decided to discard the observations where the aircraft that operated the flight is no longer part of the fleet. Only the values ‘A21N’, ‘A320’, ‘DH8B’ and ‘DH8D’ were kept in the dataset, corresponding to, respectively, the ‘Airbus A321 neo’, the ‘Airbus A320’, the ‘De Havilland Canada Dash 8-200’ and ‘De Havilland Canada Dash 8-400’.

Another critical procedure in the data preparation phase is the handling of outliers. Outliers are considered to be anomalies in the data that may be the result of incorrect measurement or failures in hardware that are inaccurate in relation to the rest of the data (Dwivedi et al. 2020). Removing outliers from a dataset is a common activity in data science projects with the purpose of discarding erroneous or noisy values (Zhao et al. 2019).

Finally, the variables “Date”, “Time” and “Flight” are important to find duplicates on the dataset, and to serve as a reference point in the case there is a need to aggregate, group or organize the data chronologically. As duplicate entries have been removed and there is no external data to be fetched and added to the current dataset, these three attributes have, at this stage, been discarded, as their presence no longer adds value to the project.

MODELLING

Once the dataset has been curated, it is ready to be used in the modelling phase. This stage is where the different Machine Learning training algorithms will be run, using the curated dataset. In some cases, different combinations of hyperparameters have been chosen within the same type of model, to be able to find the hyperparameters that yield the best result. Table 2 summarizes the results obtained for every estimator, including the Root Mean Squared Error (RMSE) and the Coefficient of Determination (R^2). It exclusively contains the metrics for each set of hyperparameters that yielded the best pair of RMSE and R^2 metrics.

Table 2 - Summary of results

	Pre-model	Linear Regression	Ridge Regression	Lasso Regression	Random Forest Regressor	AdaBoost Regressor	Neural Network
RMSE	156.99	139.31	139.32	142.98	52.84	210.39	137.74
R^2	0.75	0.80	0.80	0.79	0.97	0.54	0.80
Best parameters from grid_search	N/A	N/A	N/A	N/A	‘n_estimators’: 150	‘n_estimators’: 150	N/A

It is important to note that the airline currently uses a software to make a cargo weight prediction one hour before departure. However, this software does not learn from the variation between that prediction it makes at that time interval and the actual cargo loaded in the aircraft. Therefore, before exploring the results of the Machine Learning models, it is important to create a benchmark to compare the results with. In this thesis, the pre-model designation refers to the benchmark results that the Machine Learning models will be compared to. It is a way of generalizing the error between ‘Cargo_predicted’, which is the airline’s final cargo prediction, and ‘Cargo_actual’, which is what it has been loaded in the aircraft. This can be done by simply measuring the RMSE using the values of both variables. Using the dataset available, the RMSE for the pre-model is 156.99. This means that on average, for every prediction, the airline is making an error of 156.99kg (in excess or in deficit) when compared to the actual weight of cargo carried in a particular flight. On the other hand, the coefficient of determination (R^2) has a value of 0.75. This coefficient represents the strength of the relationship between the dependent variable (what is being predicted) and the independent variables (the features that help predict the label). The value of R^2 explains the level of proportion of the variance of the predicted values that can be explained by the predictor attributes.

A Python script, aided by both the 'sklearn' and 'tensorflow' packages, has been produced to fit the training dataset in different Machine Learning regression models. For some estimators, different combinations of hyperparameters have been inputted. This the case with the estimators 'Random Forest Regressor' and 'AdaBoost Regressor', both of which being fit with the aid of 100 or 150 Decision Trees in each run. For every fitting phase, the dataset has been split in 90% for training purposes and 10% for testing purposes. A 10-fold cross-validation split has been performed for the Linear, Ridge, Lasso, Random Forest and AdaBoost regressions.

The evaluation of the models is done by assessing the variation of both RMSE and R^2 metrics in relation to the benchmark value. Models which present a smaller RMSE and higher R^2 values perform better because the predictions calculated are closer to the true values, and the variance in the prediction can be better explained by the variables used in the training stage.

In general terms, every Machine Learning model presented in this study has produced better predictions than when compared with the pre-model value, having only one model performed worse. The results can be compared to the pre-model metrics, and consequently be split into three different groups, represented in figure 2 below.

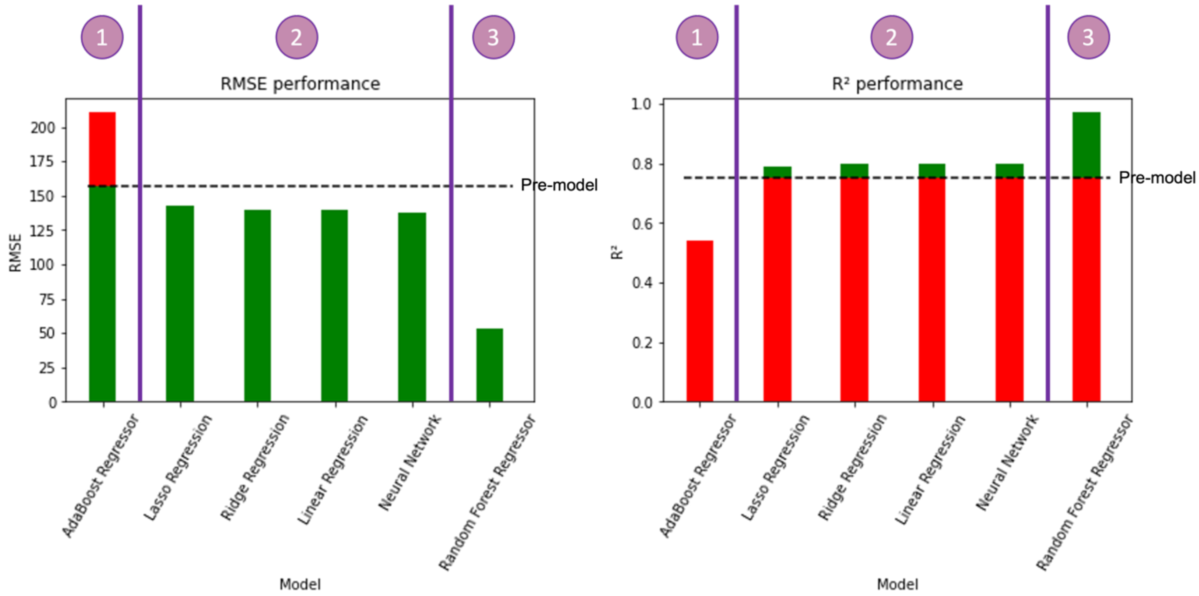


Figure 2 - A visual comparison of the Machine Learning models against the pre-model metrics.

Group 1 only contains the AdaBoost Regressor. The model in this group underperforms when compared to the pre-model. As a result, this model should not be considered by the airline to make better predictions. Group 2 contains four models that perform better than the pre-model, explained by the lower RMSE and higher R^2 values. All these models produce, in general, a more accurate prediction of the 'Cargo_actual' variable, giving the airline better forecasts. All these models, if utilized, would bring incremental value to the decision-making process. Nonetheless, it is in Group 3 that the best model is found, composed of just one estimator: the Random Forest Regression (RFR) model. This model not only outperforms the other models, but, more importantly, it is a significant improvement from its primary benchmark (the pre-model). Whilst the pre-model, on average, produces an average error of 156.99kg in its prediction, the RFR produces a lower average error of 52.84kg, which represents a 66.3%

reduction in the RMSE. The R^2 of the RFR comparatively increases from 0.75 to 0.97, a value very close to 1.0. Having an R^2 close to 1.0 indicates a good model fit and establishes a strong relationship between the variables and the prediction, where a very high proportion of the variance in the predicted variable is explained by a linear combination of the independent variables.

BUSINESS CONTEXT

Data Science is used as a mean to an end. The ultimate goal is to construct models that can produce forecasts that are relevant to scenario in which the use of data science techniques is being implemented. Along with the evaluation of the models in the context of Data Science (i.e., comparing the results among models and with a benchmark value), there is also the need to perform an assessment regarding the potential added value that the forecasts of those models bring to the business problem.

One hour before departure, the airline makes use of the software application to calculate its final cargo weight prediction, contemplated in the variable 'Cargo_prediction'. From this moment on, planners will act accordingly to ensure that the planned cargo will make its way to the aircraft. Yet, as it has been demonstrated before, this prediction does not completely correspond to the actual cargo loaded in the aircraft. This will affect the airline in many ways. Firstly, it affects the operations of the airline. Ground logistics staff guarantee that the cargo is transported between the warehouse and the aircraft and that it is loaded in the hold (and/or cabin). Any airline that aims to be competitive needs to guarantee high employees' productivity levels by optimizing its manpower planning. The staffing needs vary depending on the amount of cargo planned to be loaded in an aircraft, for each flight. As a result, a poor prediction can result in an unoptimized manpower planning. Secondly, it affects the airline’s service proposition to their cargo customers. If the airline makes an optimistic forecast for a particular flight in terms of the weight of the cargo to be transported and the actual cargo loaded in the plane is less than the predicted value, the airline will have to leave cargo behind, which can potentially lead to decreasing levels of customer satisfaction, increasing costs of rerouting and/or financial penalties for breach of contract. In the other hand, if the airline’s forecast is pessimistic, the airline could potentially carry extra cargo, to be able to maximize the output of its assets, as long as the actions required to do so are feasible in the context of the problem. Regardless of the prediction being optimistic or pessimistic, the airline needs to make sure it decreases the error between the predictions and the real values.

As a result, the airline can make use of the Machine Learning models to produce more accurate forecasts, which then result in increased efficiency in operations, better customer service and a reduction in costs. It is important to reiterate, however, the time window in which the Machine Learning models are making the predictions. One hour before departure is considered a relatively small interval to act upon a cargo prediction. Staff and resources allocation, along with the re-routing of cargo via other flights (either operated by the same airline or by a partner airline), are activities that might require more than one hour in advance to be able to be executed. Regardless of how good the model implemented is, if the actions resulting from its predictions are compromised by the time window in which the predictions are made, the usefulness of the model is being limited.

RECOMMENDATIONS

The first recommendation relates to the time window of the predictions of the models here presented. It has been mentioned before that the airline uses a software that makes predictions upon user request. This prediction can be made at any time before departure, and it constantly changes as the independent variables used also change in value. Yet, these predictions are not recorded in any database. They are for read-only purposes, to help the Reservations department in their task of maximizing the revenue from the cargo output produced by the airline. It is recommended that the airline automates the collection of the predictions of the weight of cargo, along with the values of the variables for each prediction. This way, the dataset would contain a sequential timeline of predictions, along with the other variables that help make these predictions. This sequence, if used as an input to a Machine Learning model, could be used to train new models in different time windows, as an attempt to produce accurate forecast of actual cargo loaded much more in advance than just one hour before departure. This would allow the airline to more efficiently allocate resources, manage reservations and market their cargo product.

The second recommendation has to do with the quality and reliability of the data. The airline has claimed a verification of the data before its insertion in the database, however, some incongruences have still been found in the dataset. Some of these errors are easy to detect, as they are simply not feasible in the context of the business. An example might be a negative number of adults travelling in a particular flight, or a load factor higher than 100% that cannot be explained even when all the jump seats of the aircraft are being used by staff. These events might raise doubts about the remainder of the dataset, where the veracity of the values is not as easy to verify, making it difficult to change or discard.

CONCLUSION

The aviation industry is highly competitive and dynamic, with internal and external factors changing the scenario in which airlines and airports operate. The COVID-19 has brought business opportunities for Full-service network carriers, which usually carry cargo as part of their product. The aviation industry is no exception when it comes to the potential that Data Science, and in particular, Machine Learning can have as an aid in the decision-making process. These decisions can affect various areas of the airline, such as Customer Service, Revenue Management, Reservations and Operations. Different Machine Learning models have been trained in this study, using the dataset provided by the airline, as an attempt to improve the cargo weight forecasts to be loaded in every flight. From the different models here explored, all apart from one have yielded more accurate forecasts when compared with the model currently used by the airline. Whilst most of these models have produced better forecasts, the Random Forest Regressor composed of an ensemble of 150 Decision Trees stood out due to having decreased approximately 33% the value of the root mean squared error, and by presenting a very good fit in relation to the data points due to its coefficient of determination being very close to 1.0. The forecasts produced by this model, one hour before departure, are of value for the airline, as these predictions can aid decisions, and foster better planning and asset optimization.

BIBLIOGRAPHY

Aggarwal, Divyam; Saxena, Dhish K.; Bäck, Thomas and Emmerich, Michael. 2020. Airline Crew Pairing Optimization Framework for Large Networks with Multiple Crew Bases and Hub-and-Spoke Subnetworks.

Bayrak, Tuncay. 2021. A framework for decision makers to design a business analytics platform for distributed organizations. *Techonology in society* 67:101747.

Berritella, Maria; La Franca, Luigi and Zito, Pietro. 2008. An analytical hierarchy process for ranking operating costs of low cost and full service airlines. *Journal of Air Transport Management* 15: 249-255.

Chung, Sai-Ho; Ma, Hoi-Lam; Hansen, Mark and Choi, Tsan-Ming. 2020. Data science and analytics in aviation. *Transportation Research Part E: Logistics and Transportation Review* 134:101837.

Climato, Stelvio; Gianini, Gabriele; Sepehri, Maryam; Asal, Rasool and Damiani, Ernesto. 2020. A cryptographic cloud-based approach for the mitigation of the airline cargo cancellation problem. *Journal of Information Security and Applications* 51:102462.

Dangeti, Pratap. 2017. Statistics for Machine Learning: Build supervised, unsupervised, and reinforcement learning models using both Python and R. Birmingham: Packt Publishing.

Doganis, Rigas. 2019. Flying off course. New York: Routledge.

Dwivedi, Rajendra K.; Rai, Arun K. and Kumar, Rakesh. 2020. Outlier Detection in Wireless Sensor Networks using Machine Learning Techniques: A Survey. *2020 International Conference on Electrical and Electronics Engineering (ICE3-2020)*.

Fageda, Xavier; Suau-Sanchez, Pere and Mason, Keith J.. 2015. The evolving low-cost business model: Network implications of fare bundling and connecting flights in Europe. *Journal of Air Transport Management* 42:289-296.

Forbes. 2020. Why Coronavirus May Pose More Risks To Aviation Than SARS Did. https://www.forbes.com/sites/oliverwyman/2020/01/31/why-the-wuhan-coronavirus-may-pose-more-risks-to-aviation-than-sars-did/?sh=7dc38e7a6cf2

Gottipati, Swapna; Shim, Kyong, J.; Jiang, Angeline, W. and Lee, Andre J. S. W.. 2020. Exploring the Impact of COVID-19 on Aviation Industry: A Text Mining Approach. *2020 11th IEEE Annual Information Technology, Electronics and Mobile Communication Conference (IEMCON).*

IATA. 2021. Annual Review. https://www.iata.org/contentassets/c81222d96c9a4e0bb4ff6ced0126f0bb/iata-annual-review-2021.pdf

Janić, Milan. 2019. Modeling the resilience of an airline cargo transport network affected by a large scale disruptive event. *Transportation Research Part D* 77:425-448.

Lai, Chih-Cheng; Shih, Tzu-Ping; Ko Wen-Chien; Tang Hung-Jen and Hsueh, Po-Ren. 2020. Severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) and coronavirus disease-2019 (COVID-19): The epidemic and the challenges. *International Journal of Antimicrobial Agents* 55(3).

Liu, Anyu; Kim, Yoo R. and O'Connell, John F.. 2021. COVID-19 and the aviation industry: The interrelationship between the spread of the COVID-19 pandemic and the frequency of flights on the EU market. *Annals of Tourism Research* 91: 103298.

Maertens, Sven; Pabst, Holger and Grimme, Wolfgang. 2016. The scope for low-cost connecting services in Europe — Is self-hubbing only the beginning? *Research in Transportation Business & Management* 21: 84-93.

Maneenop, Sakkakom and Kotcharin, Suntichai. 2020. The impacts of COVID-19 on the global airline industry: An event study approach. *Journal of Air Transport Management* 89:101920.

Morlotti, Chiara; Birolini, Sebastian; Cattaneo, Mattia and Redondi, Renato. 2020. Introducing connecting flights in LCCs' business model: Ryanair's network strategy. *Journal of Air Transport Management* 87.

Morrell, Peter S. and Klein, Thomas. 2018. Moving Boxes by Air 2nd Edition. London: Routledge.

Morrell, Peter, S. 2013. Airline Finance 4th edition. Farnham: Ashgate.

Niu, Baozhuang; Dai, Zhipeng and Zhuo, Xiaopo. 2019. Co-opetition effect of promised-delivery-time sensitive demand on air cargo carriers' big data investment and demand signal sharing decisions. *Transportation Research Part E: Logistics and Transportation Review* 123:29-44.

Oracle. 2021. What is Data Science? https://www.oracle.com/pt/data-science/what-is-data-science/

Özden, Aybike T. and Celik, Erkan. 2021. Analyzing the service quality priorities in cargo transportation before and during the Covid-19 outbreak. *Transport Policy* 108:34-46.

Pearce, B. 2019. Cargo outlook. https://www.iata.org/pressroom/media-kit/Documents/economics-cargo-day-2018.pdf

Ramadhani, Sameera; Dhini, Arian and Laoh, Enrico. 2020. Airline Passenger Forecasting using ARIMA and Artificial Neural Networks Approaches. International Conference on ICT for Smart Society (ICISS), ICT for Smart Society (ICISS), 2020 International Conference On CFP2013V–ART (November): 1–5.

Rizzo, Stefano G.; Chen, Yixian; Pang, Linsey; Lucas, Ji; Kaoudi, Zoi; Quiane, Jorge and Chawla, Sanjay. 2020. Prescriptive Learning for Air-Cargo Revenue Management. *2020 IEEE International Conference on Data Mining (ICDM).*

Sánchez-Rivero, Marcelino; Rodríguez-Rangel, María C. and Ricci-Risquete, Alejandro. 2021. Percepción empresarial de la pandemia por COVID-19 y su impacto en el turismo: un análisis cualitativo del destino Extremadura, España. *Journal of Management and Economics of Iberoamerica* 37(159): 265-279.

Scheiwiller, Stephanie and Zizka, Laura. 2021. Strategic responses by European airlines to the Covid-19 pandemic: A soft landing or a turbulent ride? *Journal of Air Transport Management* 95:102103.

Shaban, Ibrahim A.; Chan, F. T. S. and Chung, S. H.. 2021. A novel model to manage air cargo disruptions caused by global catastrophes such as Covid-19. *Journal of Air Transport Management* 95: 102086.

Shobande, Olatunji A. and Akinbomi, Mobolaji D.. 2020. “Competition dynamics in Nigerian aviation industry: a game theoretic approach.” *Future Business Journal* 6(1): 11.

Tsai, Chun-Wei; Chen, Yi-Ping; Tang, Tzu-Chieh and Luo Yu-Chen. 2021. An efficient parallel machine learning-based blockchain framework. *ICT Express* 7(3):300-307.

Vaast, Emmanuelle and Pinsonneault, Alain. 2021. When Digital Technologies Enable and Threaten Occupational Identity: The Delicate Balancing Act of Data Scientists. *MIS Quarterly* 45(3):1087-1112.

Vinod, Ben. 2021. Airline revenue planning and the COVID-19 pandemic. *Journal of Tourism Futures*.

Wittmer, Andreas; Bieger, Thomas and Müller, Roland. 2021. Aviation Systems Management of the Integrated Aviation Value Chain. Berlin: Springer.

Wong, Eugene Y. C.; Mo, Daniel Y. and So, Stuart. 2020. Closed-loop digital twin system for air cargo load planning operations. *International Journal of Computer Integrated Manufacturing* 34(7-8): 801-813.

Xu, Shuojiang; Chan, Hing, K. and Zhang, Tiantian. 2019. Forecasting the demand of the aviation industry using hybrid time series SARIMA-SVR approach. *Transportation Research Part E: Logistics and Transportation Review* 122:169-180.

Yaakoubi, Yassine; Soumis, François and Lacoste-Julien, Simon. 2020. Machine learning in airline crew pairing to construct initial clusters for dynamic constraint aggregation. *EURO Journal on Transportation and Logistics*.

Yang, Y., Zhang, H. and Chen, X. 2020. Coronavirus pandemic and tourism: Dynamic stochastic general equilibrium modelling of infectious disease outbreak. *Annals of Tourism Research* 83.

Yeoman, Ian S. and McMahon-Beattie, Una. 2017. The turning points of revenue management: a brief history of future evolution. *Journal of Tourism Futures* 3(1):66-72.

Zhao, Xun; Cui, Weieri; Wu, Yanhong; Zhang, Haidong; Qu, Huamin and Zhang, Dongmei. 2019. Oui! Outlier Interpretation on Multi-dimensional Data via Visual Analytics. *Eurographics Conference on Visualization (EuroVis) 2019* 38(3).

Zhou, Yaoming; Wang, Junwei and Huang, George Q. 2019. Efficiency and robustness of weighted air transport networks. *Transportation Research Part E: Logistics and Transportation Review.* 122:14-26.

APPENDIXES

In [ ]:

```python
#// IMPORTS
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt

#// READ DATA
dtypes_dic = {'Date': str, 'Time': str, 'Flight': np.int32,
              'Aircraft_type': str, 'Origin': str, 'Destination': str,
              'Adults': np.int32, 'Children': np.int32, 'Infants': np.int32,
              'Load_factor': np.float64, 'Pilot_ID': str,
              'Fish_IO': str, 'Animals_IO': str, 'METAR_origin': str,
              'Cargo_prediction': np.float64, 'Cargo_actual': np.float64}

num_attributes = [
    'Adults', 'Children', 'Infants', 'Load_factor',
    'Cargo_prediction', 'Cargo_actual'
]
cat_attributes = ['Date', 'Time',
    'Flight', 'Aircraft_type', 'Origin', 'Destination', 'Pilot_ID',
    'Fish_IO', 'Animals_IO', 'METAR_origin'
]

raw_df = pd.read_csv('./data/flight_data_raw.csv', delimiter = ';',
                  names = list(dtypes_dic.keys()), skiprows = 1,
                  dtype = dtypes_dic)

print('Raw dataset size:', raw_df.shape)
```

```
Raw dataset size: (130299, 16)
```

In [ ]:

```python
#// EXPLORATORY DATA ANALYSIS
#Create describing function
def describe_df(var_type, attributes, df):
    if var_type == 'categorical':
        print('CATEGORICAL ATTRIBUTES')
        categorical = []
        for attribute in attributes:
            missing = df[attribute].isna().sum()
            temp_series = df.loc[~df[attribute].isna(),attribute]
            unique = len(np.unique(temp_series))
            categorical.append([attribute, missing, unique])

        return pd.DataFrame(categorical, columns = ['Attribute', '# Missing valu
es',
                                                    '# Unique values'])
    elif var_type == 'numerical':
        print('NUMERICAL ATTRIBUTES')
        numerical = []
        for attribute in attributes:
            missing = df[attribute].isna().sum()
            temp_series = df.loc[~df[attribute].isna(),attribute]
            min = np.min(temp_series)
            max = np.max(temp_series)
            mean = np.mean(temp_series)
            numerical.append([attribute, missing,min, max, mean])

        return pd.DataFrame(numerical, columns = ['Attribute', '# Missing value
s', 'Min',
                                                  'Max', 'Mean'])
```

In [ ]:

```python
describe_df('categorical', cat_attributes, raw_df)
```

CATEGORICAL ATTRIBUTES

Out[ ]:

	Attribute	# Missing values	# Unique values
0	Date	0	2434
1	Time	0	190
2	Flight	0	1402
3	Aircraft_type	0	6
4	Origin	0	103
5	Destination	0	108
6	Pilot_ID	0	106
7	Fish_IO	0	2
8	Animals_IO	0	2
9	METAR_origin	123376	5668

In [ ]:

```
#// EXPLORATORY DATA ANALYSIS
describe_df('numerical', num_attributes, raw_df)
```

NUMERICAL ATTRIBUTES

Out[ ]:

	Attribute	# Missing values	Min	Max	Mean
0	Adults	0	-3.0	291.00	65.742792
1	Children	0	0.0	92.00	3.011903
2	Infants	0	0.0	13.00	0.612722
3	Load_factor	0	0.0	155.56	69.495898
4	Cargo_prediction	0	0.0	10913.00	235.074151
5	Cargo_actual	0	0.0	10969.00	298.088711

In [ ]:

```
#// DATA CURATION
df = raw_df.copy()

# Delete duplicates
df = df[~raw_df.duplicated(subset = ['Date', 'Time', 'Flight'],\
                                   keep = False)]

# Variable selection
df = df.drop(columns = ['Date', 'Time', 'Flight', 'METAR_origin'])

# Discard airliners no longer in fleet
current_fleet = ['A21N', 'A320', 'DH8B', 'DH8D']
df = df[df.Aircraft_type.isin(current_fleet)]

# Discard dummy airports
df = df[(df.Origin != 'XXX') & (df.Destination != 'XXX')]

# Discard inconsistencies in data
df = df[(df.Adults >= 0) & (df.Load_factor <= 100.0)]

#Discard disrupted flights
df = df[df.Origin != df.Destination]

# Discard outliers on 'Cargo_actual'
df1 = df[df.Aircraft_type == 'A21N']
df2 = df[df.Aircraft_type == 'A320']
df3 = df[df.Aircraft_type == 'DH8B']
df4 = df[df.Aircraft_type == 'DH8D']

def remove_outliers(df_t, list_columns):
    for column in list_columns:
        Q1 = np.quantile(df_t[column], 0.25)
        Q3 = np.quantile(df_t[column], 0.75)
        IQR = Q3-Q1
        threshold_min = Q1 - 1.5*IQR
        threshold_max = Q3 + 1.5*IQR
        df_t = df_t[(df_t[column] >= threshold_min) \
                  & (df_t[column] <= threshold_max)]
    return df_t

columns_with_outliers = [
    'Adults', 'Children', 'Infants',
    'Load_factor', 'Cargo_actual', 'Cargo_prediction'
]

df1 = remove_outliers(df1, columns_with_outliers)
df2 = remove_outliers(df2, columns_with_outliers)
df3 = remove_outliers(df3, columns_with_outliers)
df4 = remove_outliers(df4, columns_with_outliers)

df_final = pd.concat([df1, df2, df3, df4])
print('Final dataset size:', df_final.shape)
```

```
Final dataset size: (97299, 12)
```

In [ ]:

```
#// EXPORT DATASET
df_final.to_csv('./data/flight_data_curated.csv', index = False)
```

EXPLORATION

In [ ]:

```
print(np.unique(raw_df.Origin))
print(len(np.unique(raw_df.Origin)))
```

```
['ACC' 'AMS' 'BCN' 'BDA' 'BGI' 'BIO' 'BNX' 'BOD' 'BOG' 'BOS' 'BRU'
'BSL'
 'BVC' 'CAG' 'CAT' 'CDG' 'CGN' 'CHR' 'CMN' 'CPH' 'CUN' 'CVU' 'EDI'
'EMA'
 'FAO' 'FLW' 'FMO' 'FNC' 'FOR' 'FRA' 'GIG' 'GRW' 'GVA' 'HAM' 'HAV'
'HOR'
 'IBZ' 'IST' 'KEF' 'LCA' 'LCJ' 'LDE' 'LED' 'LEI' 'LGW' 'LIS' 'LOS'
'LPA'
 'LYS' 'MAD' 'MLA' 'MUC' 'MXP' 'NWI' 'OAK' 'OPO' 'ORY' 'OSR' 'OST'
'OUD'
 'OVB' 'OXB' 'PAD' 'PDL' 'PGF' 'PIX' 'PMI' 'PRN' 'PUJ' 'PVD' 'PVG'
'PXO'
 'RAI' 'RAK' 'SAW' 'SCQ' 'SID' 'SJZ' 'SKP' 'SMA' 'SSA' 'SVQ' 'TER'
'TIA'
 'TMS' 'TRN' 'TSE' 'TUS' 'VCV' 'VGO' 'VIE' 'VLC' 'VXE' 'WOE' 'XFW'
'XXX'
 'YHM' 'YMX' 'YUL' 'YYT' 'YYZ' 'ZAZ' 'ZZZ']
103
```

In [ ]:

```
print(np.unique(raw_df.Destination))
print(len(np.unique(raw_df.Destination)))
```

```
['ACC' 'AMS' 'BCN' 'BDA' 'BGI' 'BIO' 'BNX' 'BOD' 'BOG' 'BOS' 'BRU'
'BSL'
 'BVC' 'BYJ' 'CAG' 'CAT' 'CDG' 'CGN' 'CHR' 'CMN' 'CPH' 'CUN' 'CVU'
'EDI'
 'EMA' 'FAO' 'FLW' 'FMO' 'FNC' 'FOR' 'FRA' 'GIG' 'GRW' 'GVA' 'HAM'
'HAV'
 'HOR' 'IBZ' 'IST' 'KEF' 'LCA' 'LCJ' 'LDE' 'LED' 'LEI' 'LGW' 'LIS'
'LOS'
 'LPA' 'LYS' 'MAD' 'MHV' 'MLA' 'MUC' 'MXP' 'MZJ' 'NAP' 'NWI' 'OAK'
'OPO'
 'ORY' 'OSR' 'OST' 'OUD' 'OVB' 'OXB' 'PAD' 'PDL' 'PGF' 'PIX' 'PMI'
'PRN'
 'PUJ' 'PVD' 'PVG' 'PXO' 'RAI' 'RAK' 'SAW' 'SCQ' 'SID' 'SJZ' 'SKP'
'SMA'
 'SSA' 'SVQ' 'TER' 'TIA' 'TMS' 'TRN' 'TSE' 'TUS' 'VCV' 'VGO' 'VIE'
'VLC'
 'VXE' 'WOE' 'XXX' 'YHM' 'YHZ' 'YMX' 'YUL' 'YYT' 'YYZ' 'ZAG' 'ZAZ'
'ZZZ']
108
```

In [ ]:

```python
#// IMPORTS

import numpy as np
import pandas as pd
import matplotlib.pyplot as plt

from sklearn.pipeline import Pipeline
from sklearn.preprocessing import StandardScaler
from sklearn.preprocessing import OneHotEncoder
from sklearn.compose import ColumnTransformer

from sklearn.linear_model import LinearRegression
from sklearn.linear_model import Ridge
from sklearn.linear_model import Lasso
from sklearn.ensemble import RandomForestRegressor
from sklearn.ensemble import AdaBoostRegressor
from sklearn.model_selection import GridSearchCV

from sklearn.metrics import mean_squared_error
from sklearn.metrics import r2_score

#// READ DATA AND FEATURE ENGINEERING
df = pd.read_csv('./data/flight_data_curated.csv')

np.random.seed(123)

num_attributes = [
    'Adults', 'Children', 'Infants', 'Load_factor', 'Cargo_prediction'
]
cat_attributes = ['Aircraft_type', 'Origin', 'Destination', 'Pilot_ID',
    'Fish_IO', 'Animals_IO'
]

num_pipeline = Pipeline([
    ('standard_scaler', StandardScaler())
    ])

cat_pipeline = Pipeline([
    ('one_hot_encoder', OneHotEncoder(sparse = False))
    ])

print('Pre-model error')
print('RMSE:', mean_squared_error(df.Cargo_actual,\
                          df.Cargo_prediction, squared = False))
print('R^2:', r2_score(df.Cargo_actual, df.Cargo_prediction))


preprocessing = ColumnTransformer([
    ('num', num_pipeline, num_attributes),
    ('cat', cat_pipeline, cat_attributes)
    ])

X = preprocessing.fit_transform(df)
y = np.array(df.Cargo_actual)

#// MODELLING
# SKLearn package
lin_reg = LinearRegression()
ridge_reg = Ridge()
```

```
lasso_reg = Lasso()
rf_reg = RandomForestRegressor()
ada_reg = AdaBoostRegressor()

estimators = [lin_reg, ridge_reg, lasso_reg, rf_reg, ada_reg]
params = [
    {'n_jobs': [-1]}, {'solver': ['auto']}, {'max_iter': [1000]},
    {'n_estimators': [100,150], 'n_jobs': [-1]},
    {'n_estimators': [100,150]}
]


grid_list = []
for estimator_i in range(len(estimators)):
    print(str(estimators[estimator_i])[:-2])
    grid_search = GridSearchCV(estimator = estimators[estimator_i],\
                               param_grid = params[estimator_i], cv = 10)
    grid_search.fit(X, y)
    y_pred = grid_search.predict(X)
    print('RMSE:', mean_squared_error(y, y_pred, squared = False))
    print('R^2:', r2_score(y, y_pred))
    grid_list.append(grid_search)
```

```
Pre-model error
RMSE: 156.99095939371807
R^2: 0.7463145675304196
LinearRegression
RMSE: 139.31439154883708
R^2: 0.8002263487643491
Ridge
RMSE: 139.32094462699217
R^2: 0.800207554393935
Lasso
RMSE: 142.97902871988063
R^2: 0.7895781051073106
RandomForestRegressor
RMSE: 52.84767418506103
R^2: 0.9712526378644482
AdaBoostRegressor
RMSE: 210.39167436392944
R^2: 0.5443794101530876
```

In [ ]:

```
import pickle

with open('/.airline_client_best_model.pkl', 'wb') as file:
    pickle.dump(grid_list[3].estimator, file)
```

In [ ]:

```python
import tensorflow as tf
from tensorflow import keras
assert tf.__version__ >= '2.0'

# # TensorFlow package
indexes = np.arange(0, X.shape[0])
np.random.shuffle(indexes)

X_train = X[indexes[10000:]]
y_train = y[indexes[10000:]]
X_test = X[indexes[:10000]]
y_test = y[indexes[:10000]]

keras.backend.clear_session()
tf.random.set_seed(123)

hidden_layers = 20

model = keras.models.Sequential()

for _ in range(hidden_layers):
    model.add(keras.layers.Dense(100, activation = 'relu'))
model.add(keras.layers.Dense(1, kernel_initializer = 'normal'))

model.compile(loss = 'mean_squared_error',
                metrics=['mean_squared_error'])

history = model.fit(X_train, y_train, epochs = 100,
                    validation_data=(X_test, y_test))
```

```
140.6871 - mean_squared_error: 19140.6871 - val_loss: 19317.4375 - v
al_mean_squared_error: 19317.4375
Epoch 93/100
2729/2729 [==============================] - 24s 9ms/step - loss: 18
763.9203 - mean_squared_error: 18763.9203 - val_loss: 20993.9512 - v
al_mean_squared_error: 20993.9512
Epoch 94/100
2729/2729 [==============================] - 24s 9ms/step - loss: 17
805.2503 - mean_squared_error: 17805.2503 - val_loss: 22881.0469 - v
al_mean_squared_error: 22881.0469
Epoch 95/100
2729/2729 [==============================] - 24s 9ms/step - loss: 55
037.9618 - mean_squared_error: 55037.9618 - val_loss: 19746.9453 - v
al_mean_squared_error: 19746.9453
Epoch 96/100
2729/2729 [==============================] - 24s 9ms/step - loss: 24
106.0743 - mean_squared_error: 24106.0743 - val_loss: 19038.5918 - v
al_mean_squared_error: 19038.5918
Epoch 97/100
2729/2729 [==============================] - 24s 9ms/step - loss: 95
29153.5662 - mean_squared_error: 9529153.5662 - val_loss: 19787.5430
- val_mean_squared_error: 19787.5430
Epoch 98/100
2729/2729 [==============================] - 24s 9ms/step - loss: 18
538.8770 - mean_squared_error: 18538.8770 - val_loss: 18819.9355 - v
al_mean_squared_error: 18819.9355
Epoch 99/100
2729/2729 [==============================] - 24s 9ms/step - loss: 70
4283.3332 - mean_squared_error: 704283.3332 - val_loss: 18556.1113 -
val_mean_squared_error: 18556.1113
Epoch 100/100
2729/2729 [==============================] - 24s 9ms/step - loss: 14
91244.4463 - mean_squared_error: 1491244.4463 - val_loss: 18972.1016
- val_mean_squared_error: 18972.1016
```

In [ ]:

```
y_pred_nn = model.predict(X_test)

print('NeuralNetwork')
print('RMSE:', mean_squared_error(y_test, y_pred_nn, squared = False))
print('R²:', r2_score(y_test, y_pred_nn))
```

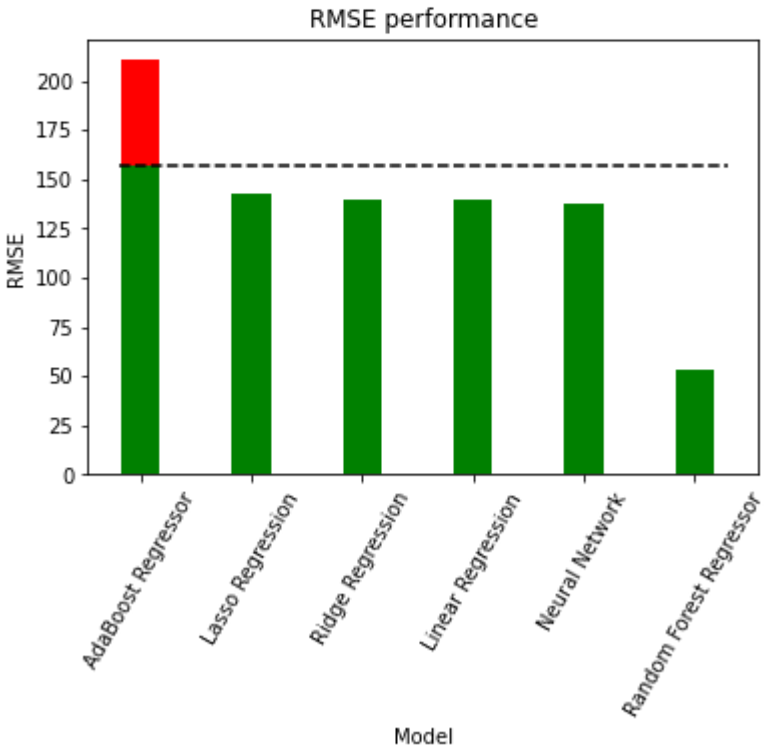
```
NeuralNetwork
RMSE: 137.73924490447317
R²: 0.8025998469615319
```

In [ ]:

```
threshold = 156.99
values = np.array([210.39, 142.98, 139.32, 139.31, 137.74, 52.84])
x = np.array(['AdaBoost Regressor', 'Lasso Regression', 'Ridge Regression', 'Lin
ear Regression', 'Neural Network', 'Random Forest Regressor'])

# split it up
above_threshold = np.maximum(values - threshold, 0)
below_threshold = np.minimum(values, threshold)

# and plot it
fig, ax = plt.subplots()
ax.bar(x, below_threshold, 0.35, color="g")
ax.bar(x, above_threshold, 0.35, color="r",
        bottom=below_threshold)

# horizontal line indicating the threshold
ax.plot([-0.2, 5.3], [threshold, threshold], "k--")
ax.set_xticklabels(x, rotation=60)
ax.set_title('RMSE performance')
ax.set_ylabel('RMSE')
ax.set_xlabel('Model')
```

Out[ ]:

```
Text(0.5, 0, 'Model')
```



In [ ]:

```
threshold = 0.75
values = np.array([0.54, 0.79, 0.80, 0.80, 0.80, 0.97])
x = np.array(['AdaBoost Regressor', 'Lasso Regression', 'Ridge Regression', 'Lin
ear Regression', 'Neural Network', 'Random Forest Regressor'])

# split it up
above_threshold = np.maximum(values - threshold, 0)
below_threshold = np.minimum(values, threshold)

# and plot it
fig, ax = plt.subplots()
ax.bar(x, below_threshold, 0.35, color="r")
ax.bar(x, above_threshold, 0.35, color="g",
        bottom=below_threshold)

# horizontal line indicating the threshold
ax.plot([-0.2, 5.3], [threshold, threshold], "k--")
ax.set_xticklabels(x, rotation=60)
ax.set_title('R² performance')
ax.set_ylabel('R²')
ax.set_xlabel('Model')
```

Out[ ]:

```
Text(0.5, 0, 'Model')
```