

Project Proposal

How do we predict flight delays at take-off?

in this notebook, I develop a model aimed at predicting flight delays at take-off. The purpose is not to obtain the best possible prediction but rather to emphasize on the various steps needed to build such a model. Along this path, I then put in evidence some **basic but important** concepts. Among then, I comment on the importance of the separation of the dataset during the training stage and how **cross-validation** helps in determining accurate model parameters. I show how to build **linear** and **polynomial** models for **univariate** or **multivariate regressions** and also, I give some insight on the reason why **regularisation** helps us in developing models that generalize well.

Dataset:

In the study of predicting flight delays, we used a dataset on Kaggle.

The dataset is available as the .csv file.

The dataset is available as the .csv file. a sample of data is shown in the following table:

	Year	Month	DayofMonth	DayOfWeek	DepTime	CRSDepTime	ArrTime	CRSArrTime	FlightNum	ActualElapsed1
count	123523.000000	123523.000000	123523.000000	123523.000000	121172.000000	123523.000000	120867.000000	123523.000000	123523.000000	120866.000000
mean	1998.624572	6.554366	15.76980	3.938513	1349.497095	1335.154886	1493.076812	1491.198724	1364.833634	120.230000
std	6.226778	3.444210	8.77794	1.984338	479.424132	477.991652	500.833239	494.282101	1405.703898	68.560000
min	1987.000000	1.000000	1.000000	1.000000	1.000000	0.000000	1.000000	0.000000	1.000000	11.000000
25%	1993.000000	4.000000	8.000000	2.000000	932.000000	926.000000	1115.000000	1115.000000	450.000000	70.000000
50%	1999.000000	7.000000	16.000000	4.000000	1335.000000	1330.000000	1522.000000	1520.000000	943.000000	102.000000
75%	2004.000000	10.000000	23.000000	6.000000	1740.000000	1730.000000	1918.000000	1914.000000	1709.000000	151.000000
max	2008.000000	12.000000	31.000000	7.000000	2505.000000	2359.000000	2608.000000	2400.000000	9581.000000	1650.000000

8 rows × 24 columns

Tools:

There are tools that will be used to achieve the goal of this study, such as:

TensorFlow

matplotlib

pandas

nltk

TO DO:

Explore the data and come up with EDA phases then use a model to fit the data.

NOTE: the used features may be increased or changed and the model as well