Machine Learning Projects (SC)

The objective of the projects is to prepare you to apply different machine learning algorithms to real-world tasks. This will help you to increase your knowledge about the workflow of the machine learning tasks. You will learn how to clean your data, applying pre-processing, feature engineering, regression, and classification methods. Each project will be delivered in milestones.

- The best three teams for each project will be honored.
- Team and Projects' Registration **starts**: Thursday 24/3/2022 11:00PM.
- Registration **ends**: Tuesday 5/4/2022 11:59PM.
- Delivering Milestone 1: 21/4/2022.
- Delivering Milestone 2: Practical exam.
- Minimum number of members is 3 and the maximum is 5
- You must deliver a detailed report for each milestone contains all your work (feature analysis, algorithms used in each module and the achieved accuracy for each one)

Note : Each report will be graded

In the first milestone, you will apply the following:-

Preprocessing: Before building your models, you need to make sure that the dataset is clean and ready-to-use.

Regression: Apply different regression techniques (at least two) to find the model that fits your data with minimum error.

Milestone 1:

- Preprocessing, Regression.

Milestone 1 Report Must Include:

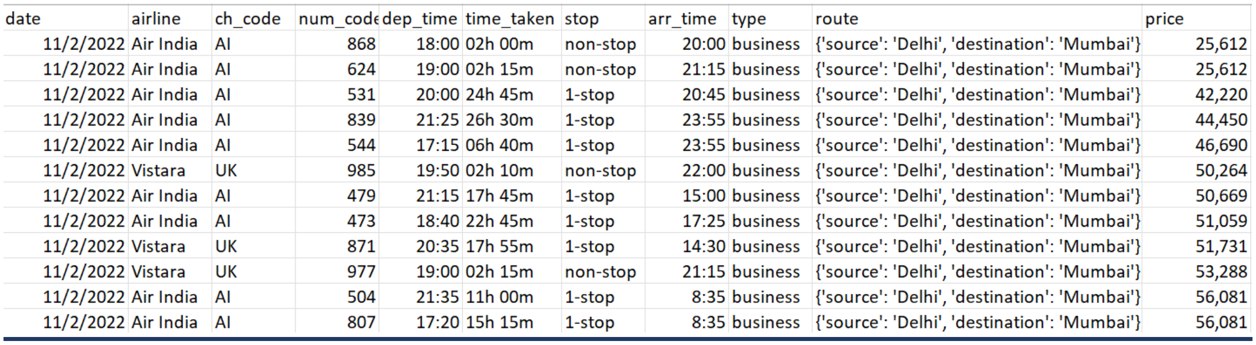
- You must explain in details the **preprocessing techniques** you needed to apply on your dataset and how you implemented them.
- Perform **analysis** on the dataset as studied and explain how the features affect and relate to each other.
- You must explain what **regression techniques** you used (at least two).
- Mention the **differences** between each model and the acquired **results** (accuracy/error and so on) and the **training time** for each model.
- You must clearly mention **what features** you used or discarded to create your regression models.
- Explain what the **sizes** of your training, testing and validation sets are, if exist.
- Mention any further techniques that were used to **improve** the results (if exist).
- You should include **screenshots** of the resultant(s) regression line plots if possible or any data visualization.
- Finally, write a **conclusion** about this phase of the project and what intuition you had about your problem and how it was proved/disproved.

Milestone 2 Deliverables will be announced later.

Project(1): Airline Ticket Price Prediction

Airline ticket pricing changes according to a number of factors such as the type of ticket, the flight time and more. Given this dataset, we want to understand which factors affect the ticket pricing the most and be able to predict future flight prices.

Dataset Snapshot:



date	airline	ch_code	num_code	dep_time	time_taken	stop	arr_time	type	route	price
11/2/2022	Air India	AI	868	18:00	02h 00m	non-stop	20:00	business	{'source': 'Delhi', 'destination': 'Mumbai'}	25,612
11/2/2022	Air India	AI	624	19:00	02h 15m	non-stop	21:15	business	{'source': 'Delhi', 'destination': 'Mumbai'}	25,612
11/2/2022	Air India	AI	531	20:00	24h 45m	1-stop	20:45	business	{'source': 'Delhi', 'destination': 'Mumbai'}	42,220
11/2/2022	Air India	AI	839	21:25	26h 30m	1-stop	23:55	business	{'source': 'Delhi', 'destination': 'Mumbai'}	44,450
11/2/2022	Air India	AI	544	17:15	06h 40m	1-stop	23:55	business	{'source': 'Delhi', 'destination': 'Mumbai'}	46,690
11/2/2022	Vistara	UK	985	19:50	02h 10m	non-stop	22:00	business	{'source': 'Delhi', 'destination': 'Mumbai'}	50,264
11/2/2022	Air India	AI	479	21:15	17h 45m	1-stop	15:00	business	{'source': 'Delhi', 'destination': 'Mumbai'}	50,669
11/2/2022	Air India	AI	473	18:40	22h 45m	1-stop	17:25	business	{'source': 'Delhi', 'destination': 'Mumbai'}	51,059
11/2/2022	Vistara	UK	871	20:35	17h 55m	1-stop	14:30	business	{'source': 'Delhi', 'destination': 'Mumbai'}	51,731
11/2/2022	Vistara	UK	977	19:00	02h 15m	non-stop	21:15	business	{'source': 'Delhi', 'destination': 'Mumbai'}	53,288
11/2/2022	Air India	AI	504	21:35	11h 00m	1-stop	8:35	business	{'source': 'Delhi', 'destination': 'Mumbai'}	56,081
11/2/2022	Air India	AI	807	17:20	15h 15m	1-stop	8:35	business	{'source': 'Delhi', 'destination': 'Mumbai'}	56,081

Milestone 1 tasks:

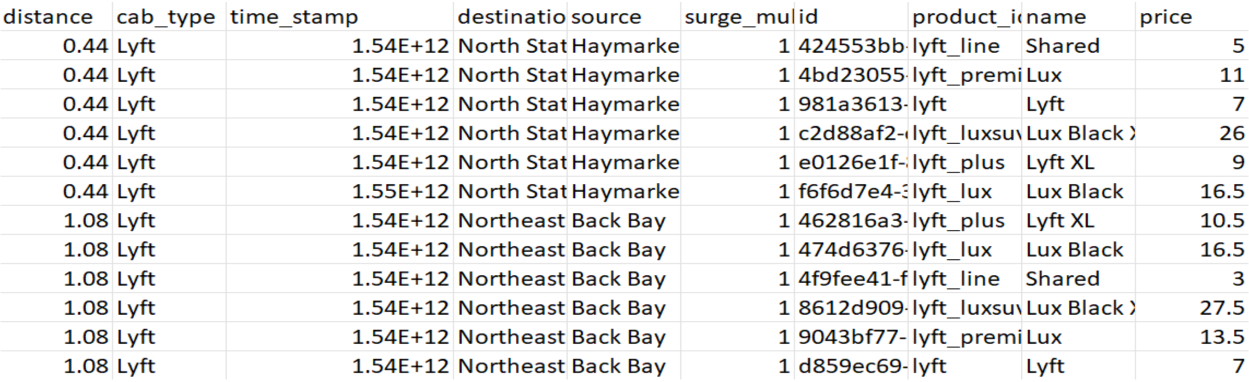
1. Apply pre-processing on the provided dataset. (You must preprocess all the features even if you won’t use them later after feature selection)
2. Apply Feature Selection and Experiment with regression techniques to reduce the error on prediction of the price of a ticket (Deliver at least two regression models with significant difference).
3. Finish Milestone 1 Report.

Note: You must preprocess all features, but model and feature selection can be done after that (i.e You can drop a feature only after preprocessing and with valid reason)

Project(2): Taxi Service Price Prediction

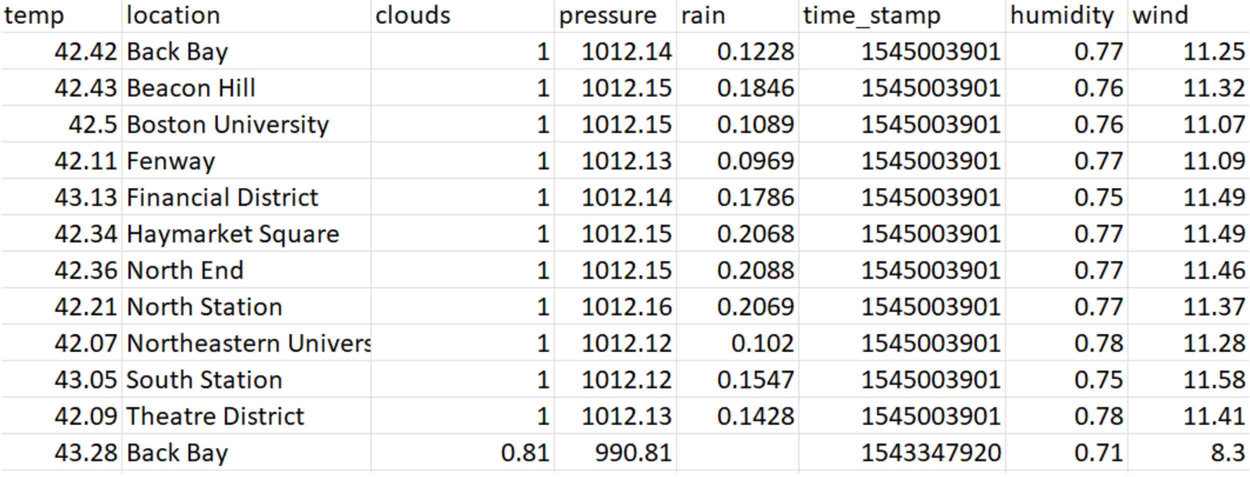
The use of taxi service providers such as Uber, Kareem and Lyft has become almost essential in recent years. Each company has their own methods of pricing each ride. These prices may be affected by the locations or the weather. Given this dataset, our task is to predict the price of a taxi ride based on the provided information.

Dataset Snapshots – File 1 (taxi_rides.csv):



distance	cab_type	time_stamp	destinatio	source	surge_mul	id	product_i	name	price
0.44	Lyft	1.54E+12	North Stat	Haymarke	1	424553bb-	lyft_line	Shared	5
0.44	Lyft	1.54E+12	North Stat	Haymarke	1	4bd23055-	lyft_premi	Lux	11
0.44	Lyft	1.54E+12	North Stat	Haymarke	1	981a3613-	lyft	Lyft	7
0.44	Lyft	1.54E+12	North Stat	Haymarke	1	c2d88af2-	lyft_luxsu	Lux Black X	26
0.44	Lyft	1.54E+12	North Stat	Haymarke	1	e0126e1f-	lyft_plus	Lyft XL	9
0.44	Lyft	1.55E+12	North Stat	Haymarke	1	f6f6d7e4-3	lyft_lux	Lux Black	16.5
1.08	Lyft	1.54E+12	Northeast	Back Bay	1	462816a3-	lyft_plus	Lyft XL	10.5
1.08	Lyft	1.54E+12	Northeast	Back Bay	1	474d6376-	lyft_lux	Lux Black	16.5
1.08	Lyft	1.54E+12	Northeast	Back Bay	1	4f9fee41-f	lyft_line	Shared	3
1.08	Lyft	1.54E+12	Northeast	Back Bay	1	8612d909-	lyft_luxsu	Lux Black X	27.5
1.08	Lyft	1.54E+12	Northeast	Back Bay	1	9043bf77-	lyft_premi	Lux	13.5
1.08	Lyft	1.54E+12	Northeast	Back Bay	1	d859ec69-	lyft	Lyft	7

Dataset Snapshots – File 2 (weather.csv):



temp	location	clouds	pressure	rain	time_stamp	humidity	wind
42.42	Back Bay	1	1012.14	0.1228	1545003901	0.77	11.25
42.43	Beacon Hill	1	1012.15	0.1846	1545003901	0.76	11.32
42.5	Boston University	1	1012.15	0.1089	1545003901	0.76	11.07
42.11	Fenway	1	1012.13	0.0969	1545003901	0.77	11.09
43.13	Financial District	1	1012.14	0.1786	1545003901	0.75	11.49
42.34	Haymarket Square	1	1012.15	0.2068	1545003901	0.77	11.49
42.36	North End	1	1012.15	0.2088	1545003901	0.77	11.46
42.21	North Station	1	1012.16	0.2069	1545003901	0.77	11.37
42.07	Northeastern Univers	1	1012.12	0.102	1545003901	0.78	11.28
43.05	South Station	1	1012.12	0.1547	1545003901	0.75	11.58
42.09	Theatre District	1	1012.13	0.1428	1545003901	0.78	11.41
43.28	Back Bay	0.81	990.81		1543347920	0.71	8.3

Milestone 1 tasks:

1. Apply pre-processing on the provided dataset. (You must preprocess the features provided in both files and add the information in the second file to the information in the first file in a meaningful way)
2. Apply Feature Selection and Experiment with regression techniques to reduce the error on prediction of the price of a taxi ride (Deliver at least two regression models with significant difference).
3. Finish Milestone 1 Report.

   Note: You must preprocess all features, but model and feature selection can be done after that (i.e You can drop a feature only after preprocessing and with valid reason)