

MADRID

Final Project — Startup valuation estimator

Alvaro Garcia Blazquez

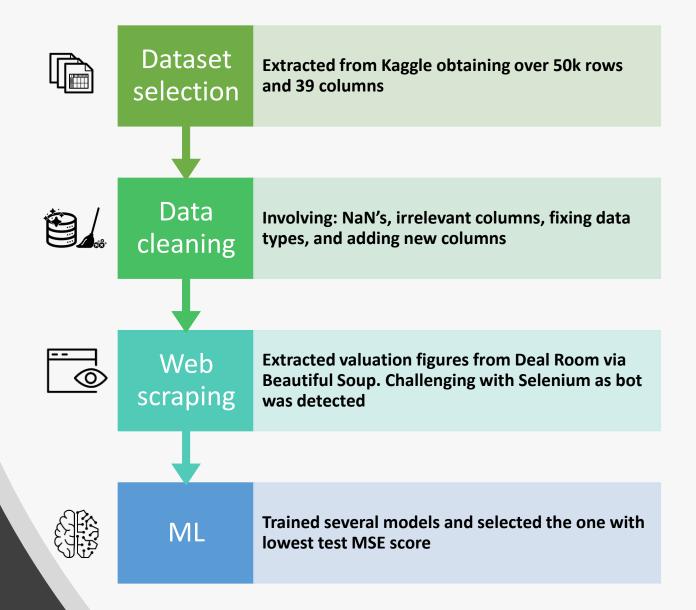


## Project Summary

- **Purpose:** Provide an estimated valuation for a startup based on key parameters inputted by the user (investor, VC, founder, individual)
- Problem to support: When determining premoney valuations in startups financing rounds, several conflicts tend to appear between founders and investors. Goal is to provide an objective / neutral approach based purely on data
- Dataset source: Kaggle
- Key tools used:
  - a) Pandas: Data exploration and cleaning
  - b) Web scraping: Dataset enrichment
  - c) ML: Model training, selection, and prediction
  - d) Streamlit: App/Web creation



## Key steps





## Issues through the process

#### Dataset selection:

 Finding data as recent as possible to take into account new startups and funding rounds

### Data cleaning:

- Criteria for eliminating rows with NaN's
- Incorrect data types

### Web scraping:

• Initially to be done with Selenium but bot was detected. Beautiful Soup was time consuming

#### ML:

- Decision of target variable (operating status vs valuation)
- Once target decided, avoiding bias towards larger predictions. Use of log(10) worked out



## ML

### **Models tried**

- Logistic Regression with **operating status** as target
- Simple Linear Regression with **absolute valuation** as target
- Linear Regression, Ridge, Lasso, SGDRegressor, KNeighborsRegressor, and GradientBoostingRegressor with log(10) of absolute valuation as target
- Decision Trees and random forest iteration via GridSearch with log(10) of absolute valuation as target

## **Predictor variables used for training**

- Total funding
- Number of funding rounds
- Year founded
- Delta between year founded and first financing
- Country\_ranking (Rank encoding of country variable)
- Merket\_ranking (Rank encoding of market variable)



# ML – Selection process

## Grid Search for Decision Trees:

```
gs = GridSearchCV(
    estimator=DTR(),
    param_grid={
        "max_depth": [5, 6],
        "min_samples_split": [50, 100, 300, 1000],
        "max_features": [4, 6]
    },
    cv=5,
    verbose=3,
    scoring="neg_mean_squared_error",
    return_train_score=True
)
```

## Decision Trees ranking based on MSE:

	param_max_depth	param_max_features	param_min_samples_split	mean_test_score	mean_train_score
14	6	6	300	-0.428390	-0.410767
13	6	6	100	-0.428608	-0.404567
5	5	6	100	-0.432529	-0.416711
12	6	6	50	-0.433882	-0.398939
6	5	6	300	-0.434674	-0.419935
4	5	6	50	-0.436700	-0.413155
15	6	6	1000	-0.442364	-0.433387
7	5	6	1000	-0.446128	-0.437846
10	6	4	300	-0.446377	-0.427722
9	6	4	100	-0.448681	-0.420704



## Potential next steps



Predicting operating status and then based on that a valuation if the prediction is that the startup will be operating



Different currencies available for total funding, and predicted valuation



Categorize "markets" variable into broader / more general blocks



Further ML training with different variables obtained via web scraping (founders education, employees, users, app downloads...)



Let's try it!

