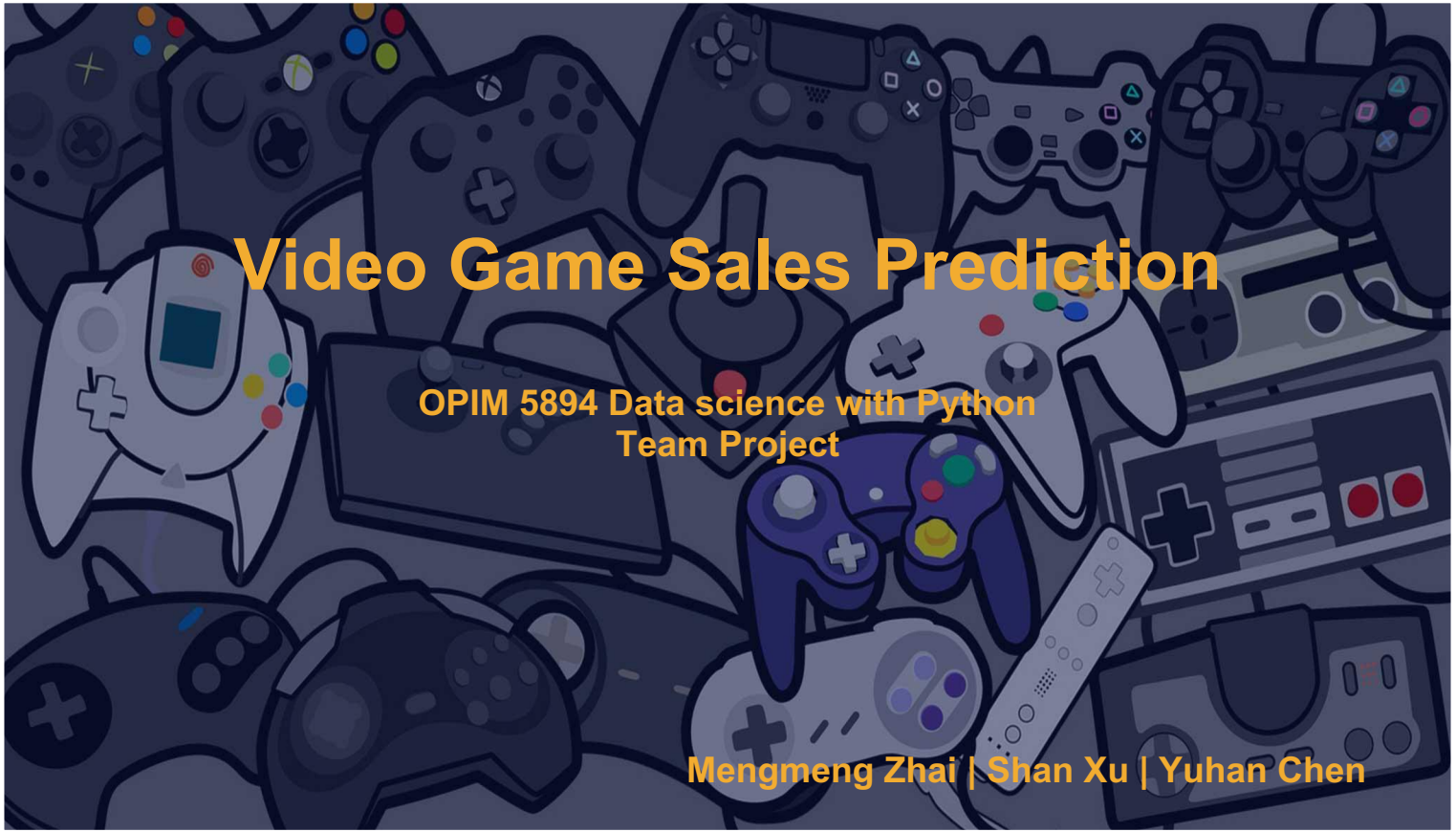


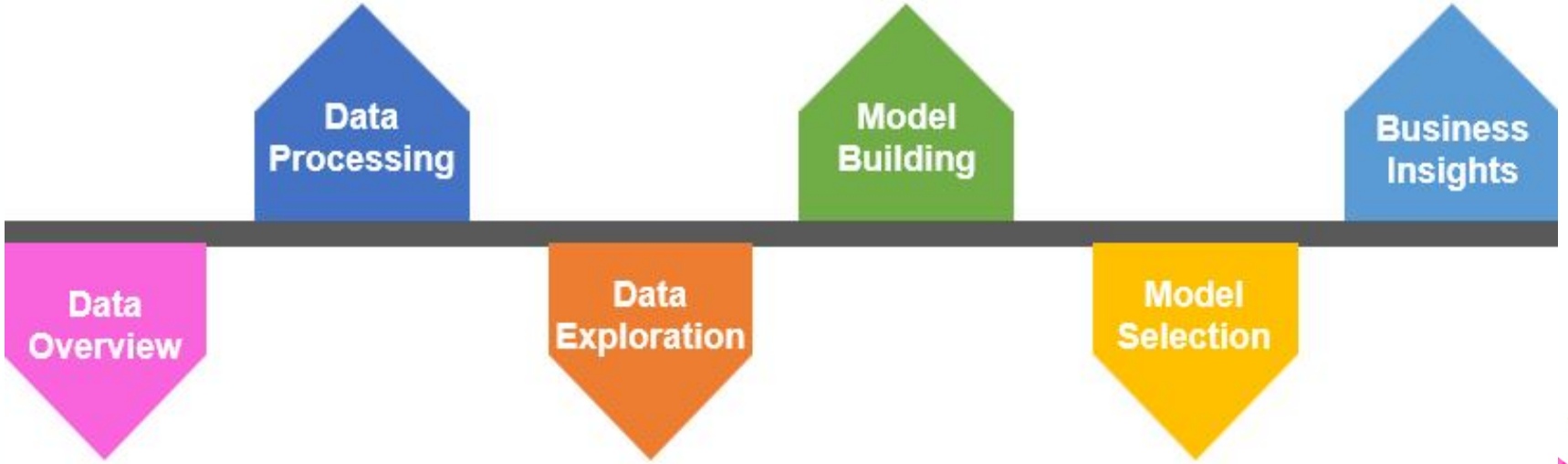
Video Game Sales Prediction

OPIM 5894 Data science with Python
Team Project

Mengmeng Zhai | Shan Xu | Yuhan Chen

Project Overview

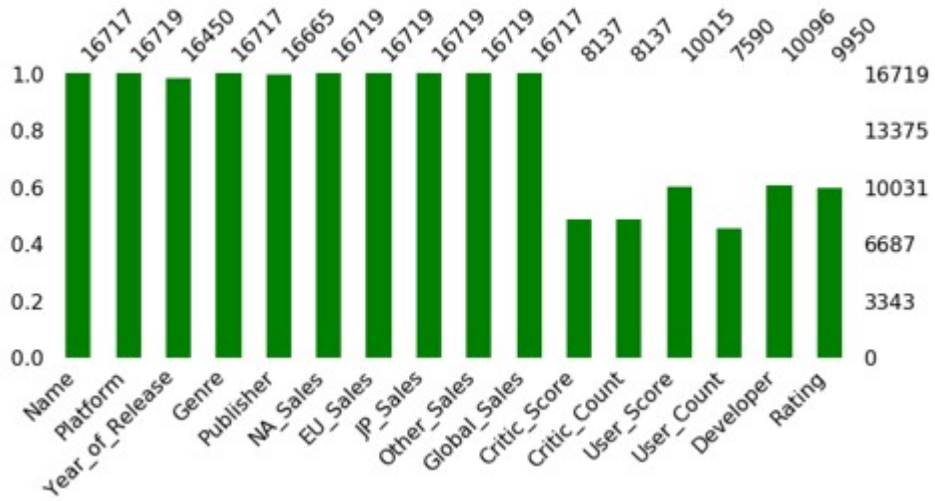
- Project Motivation
- Project Workflow



Data Overview

- The dataset obtained is mainly about regional sales(EU/NA/JP/Other) and scores of video games.
- The shape of the original dataset: (16719, 16)
- 6 categorical variables
- Challenges:

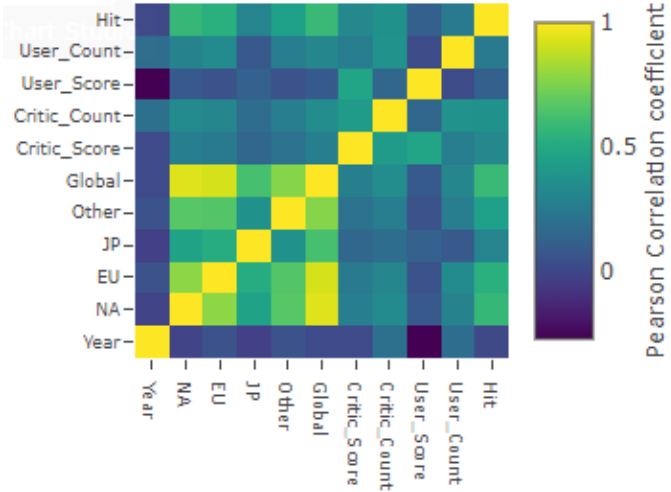
  -missing value in some columns (Developer/ Rating/ User_score/ User_count/ Critic_score/ Critic_count)
  -too many levels in some categorical variables (Publisher / Developer/ Name)



	# of Missing Values	% of Total Values
User_Count	9129	54.6
Critic_Score	8582	51.3
Critic_Count	8582	51.3
Rating	6769	40.5
User_Score	6704	40.1
Developer	6623	39.6
Year_of_Release	269	1.6
Publisher	54	0.3
Name	2	0.0
Genre	2	0.0
Global_Sales	2	0.0

Data Processing

- Deal with missing values
  - -drop observations if ‘Developer/Rating/User_Score/User_Count/Critic_Score/Critic_Count’ are blank
  - -replace nan with the median
- Keep more frequent dummy variables for ‘Name/ Publisher/ Developer’
- Correlation between variables
  - -Global sales **vs.** JP/NA/EU/Other sales
- Create binary column ’Hit’ based on ‘Global_Sales’ for further use
  - -If global sale >$1 million, Hit is 1; otherwise Hit is 0
- Use up-sample method to handle with imbalanced class problem




```
df1.Hit.value_counts()
0    7158
1    1393
```

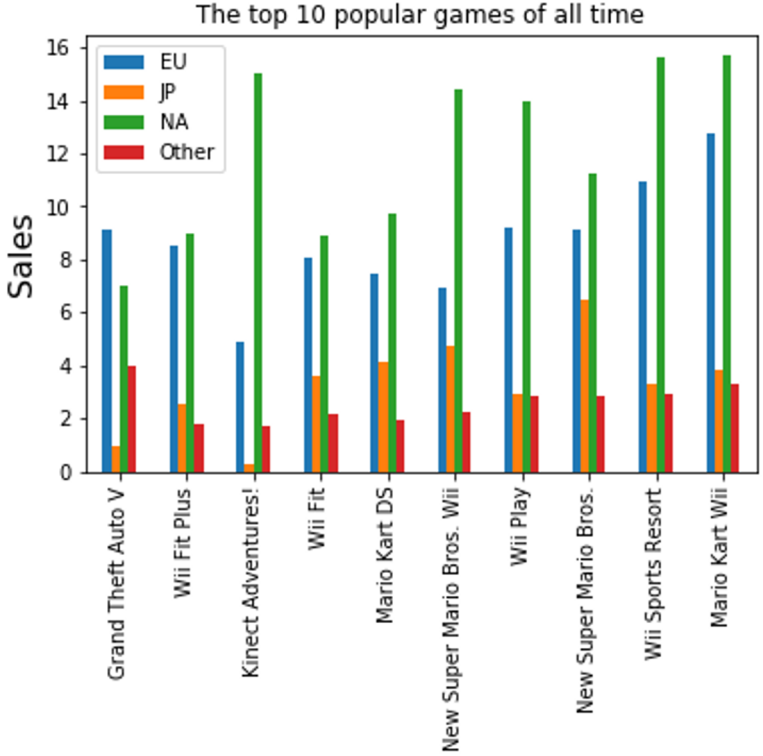
```
df1.Hit.value_counts()
1    7158
0    7158
```

Objective Statement

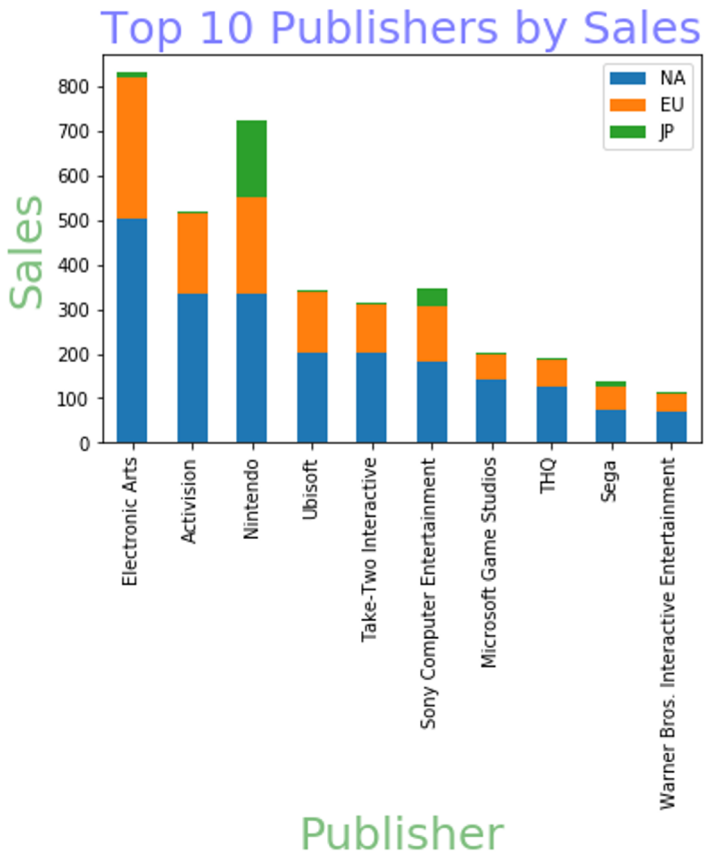
Regression → Predict the global sales

Classification → Predict the popularity (Hit)

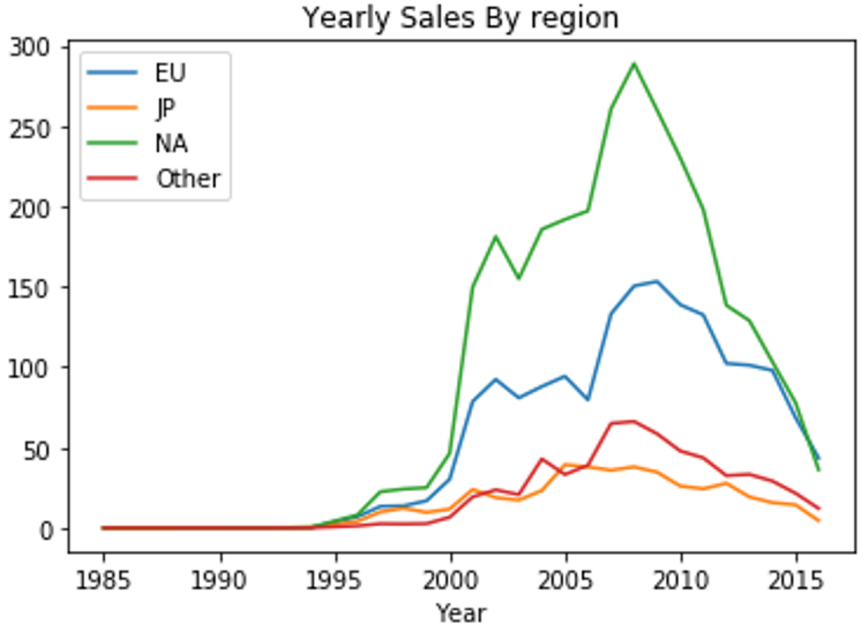
Data Exploration

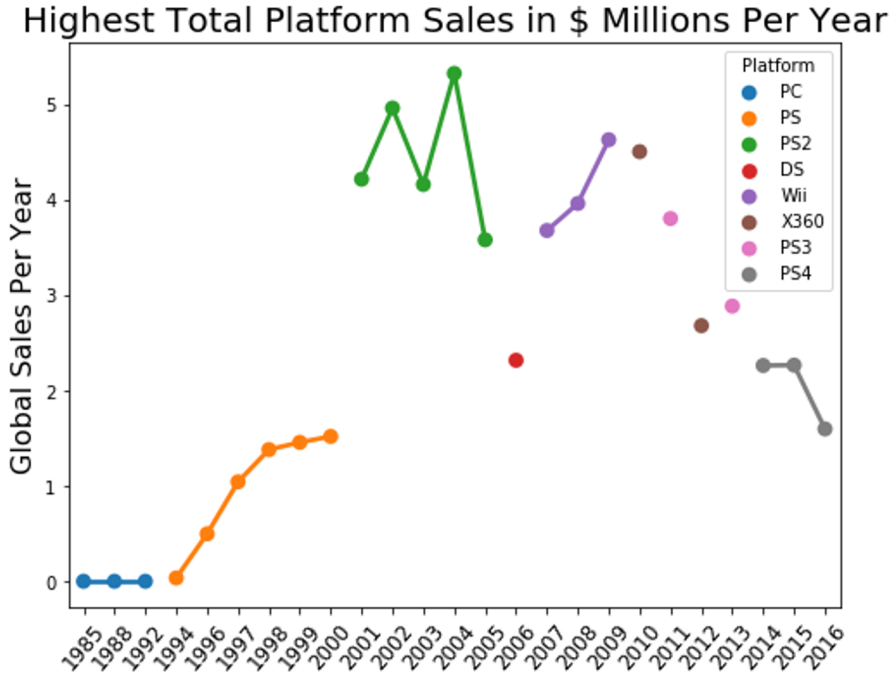


Data Exploration - continued



Data Exploration





Model Building

- What we learned in class and more
- 5 models for regression and 4 models for classification
- Regression: Linear regression, kNN, Random Forest, Gradient Boosting, Neural Network
- Classification: Logistic Regression, Decision Tree, SVM, Gaussian Naïve Bayes

Model Building

- Both are ensemble machine learning methods
- GB build trees one at a time, where each new tree helps to correct errors made by previously trained tree.
- RF trains each tree independently, using a random sample of the data.
- Weakness comparison

Gradient Boosting	Random Forest
➢ longer training time; ➢ hard to tune; ➢ sensitive to overfitting if data is noisy	➢ slow for real-time prediction; ➢ biased in favor of attributes (from categorical variables)

Model Building

- Decision Tree (criterion='entropy', max_depth=3)

- Logistic Regression:
  Default parameters (penalty='l2', C=1.0, fit_intercept=True)
  C: Regularization strength
  fit_intercept: compute intercept of linear classifier or not
  penalty: whether to use Lasso or Ridge Regularization

- Neural Network
  Use Sequential model, which is a linear stack of layers: 3 hidden layers with Dense(32, activation='relu'), 1 output layer

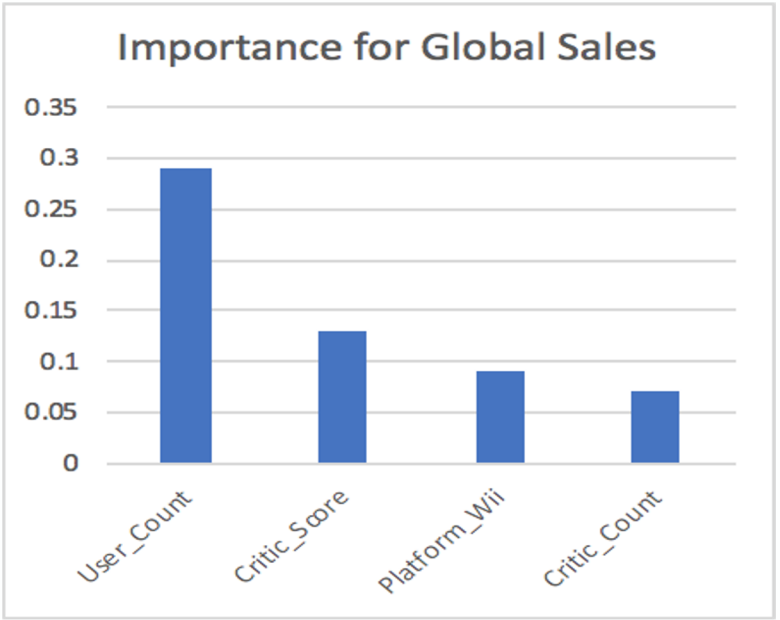
Model Selection-Regression

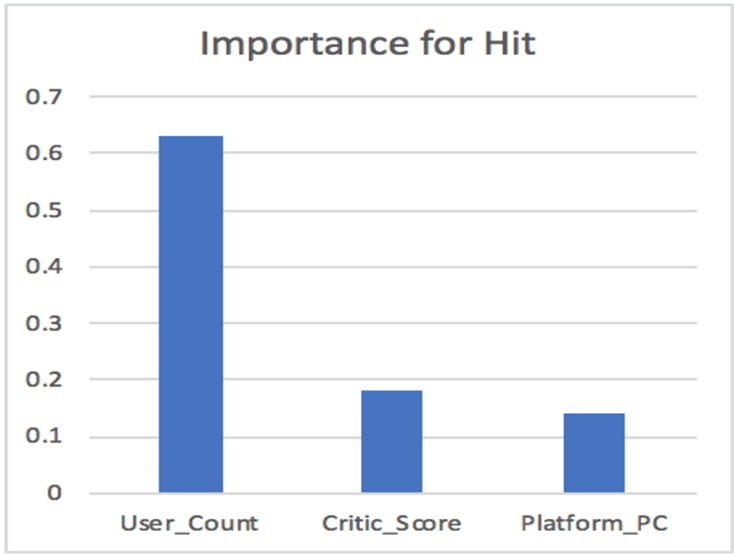
	RMSE	MAE	Explain Variance Score	R squared
Gradient Boosting	0.00088	0.0177	0.7994	0.7992
Neural Network	0.0126	0.0896	0.2279	0.2366
Random Forest	0.00853	0.0497	0.8238	0.9536
kNN	0.00895	0.0253	0.9856	0.9850
Linear Regression	0.0570	0.0302	0.2632	0.2626

Model Selection-Classification

	Accuracy score	Recall score	Precision score	AUROC
Logistic Regression	0.7667	0.7602	0.7709	0.7667
Decision Tree	0.7739	0.8008	0.7654	0.7734
Support Vector Machine	0.7343	0.6957	0.7582	0.7347
Gaussian Naïve Bayes	0.5804	0.4741	0.5246	0.5110

Finding-Variable Significance





Literature and Distinction

- **Literature**

  Visualization with box plots and scatter plot
  Classification of Popularity
  Global Sales Prediction

- **Distinction**

  Both regression & classification
  Up-sample method for imbalanced class
  Only keep some dummy variables to decrease model complexity

Business Insights

- The video games published by Electronic Arts have higher business returns.
- The video games in Wii platform tend to have higher amount of business sales whereas the games in PC platform tend to be more popular.
- The number of users who give the score at Metacritic and the aggregate score compiled by Metacritic's staff can indicate whether the market prospect is positive for that video game.
- Decrease the weight of the game genre when considering the game popularity and the sales from business perspective.

Business Insights

- Employ the gradient boosting model for marketing and sales departments.
- Employ the kNN model for internal statistical research or further model enhancement.

Global Sales

- Employ the decision tree model when predicting the popularity of games regionally or globally.
- Additional factors may need to be considered to increase the prediction power, such as the investment, music, and etc.

Hit

Reference

- https://stackabuse.com/implementing-svm-and-kernel-svm-with-pythons-scikit-learn/
- https://scikit-learn.org/stable/modules/generated/sklearn.preprocessing.StandardScaler.html
- https://datascienceplus.com/extreme-gradient-boosting-with-python/
- https://stackoverflow.com/questions/49008074/how-to-create-a-neural-network-with-regression-model
- https://towardsdatascience.com/random-forest-in-python-24d0893d51c0


Thank you!