**CS-514**

**Applied Artificial Intelligence**

**Project**

[DATASET: House Prices using Advanced Regression Techniques](https://www.kaggle.com/c/house-prices-advanced-regression-techniques/overview)

**Instructions to run:**

* Requirements: Python3, Jupyter, Anaconda (numpy, pandas, sklearn, scipy, matplotlib, xgboost)
* Download the file and unzip to find notebook.ipynb
* Start Jupyter Notebook and open the file in browser
* Run the file
* The output files for each model are created in the data folder

**Data Pre-Processing:**

**Dealing with outliers:**

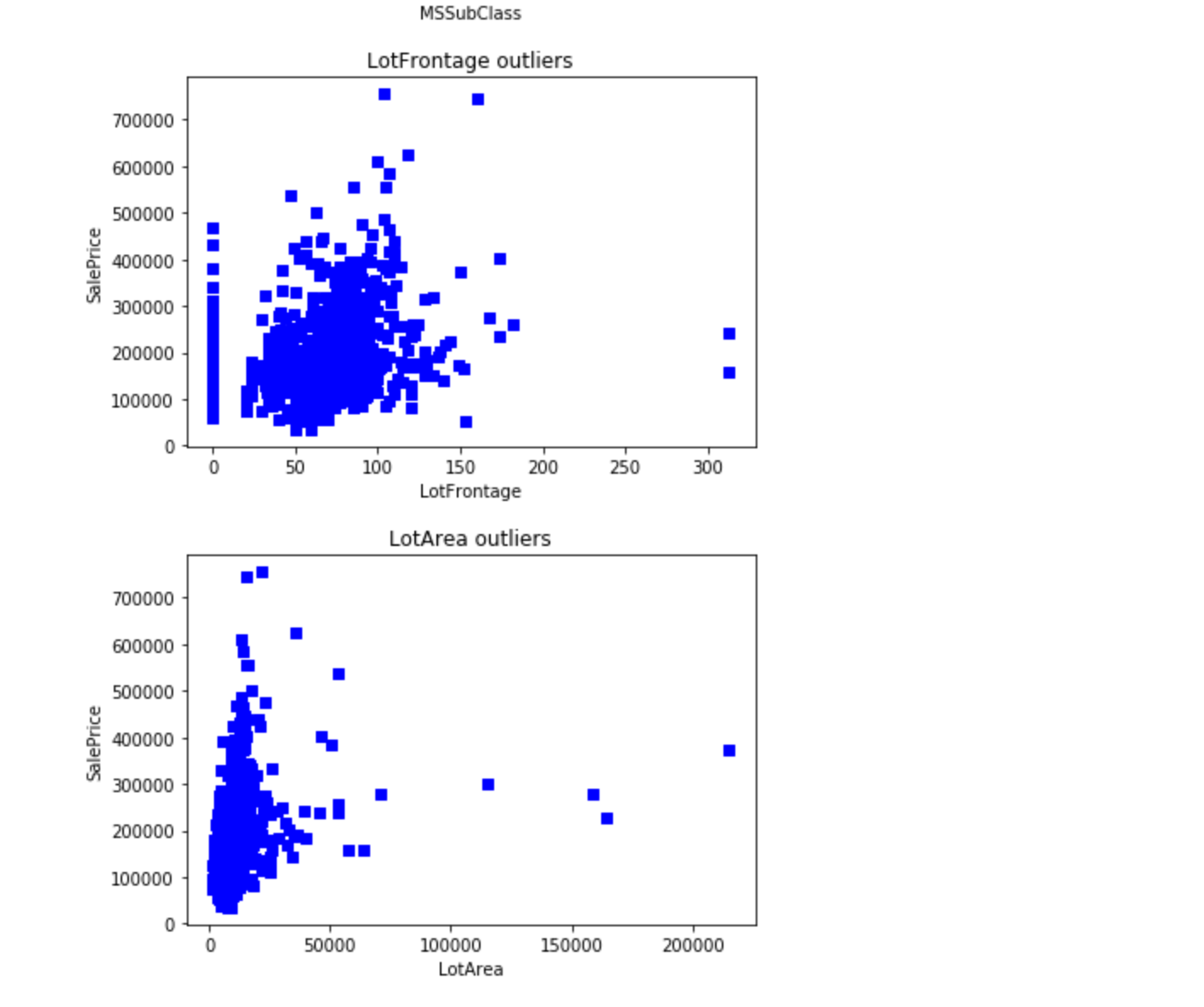
I run through all the numerical features present in the data and plot the values on a scatter plot to determine what outliers are present.

Once we look at all the features and realize their outliers, we then restrict the dataset to remove outliers and only consider data that won’t skew the model.

While doing so, it may so happen that if we remove all outliers the number of records can become very small, so I check if the dataset has enough number of records for modelling.

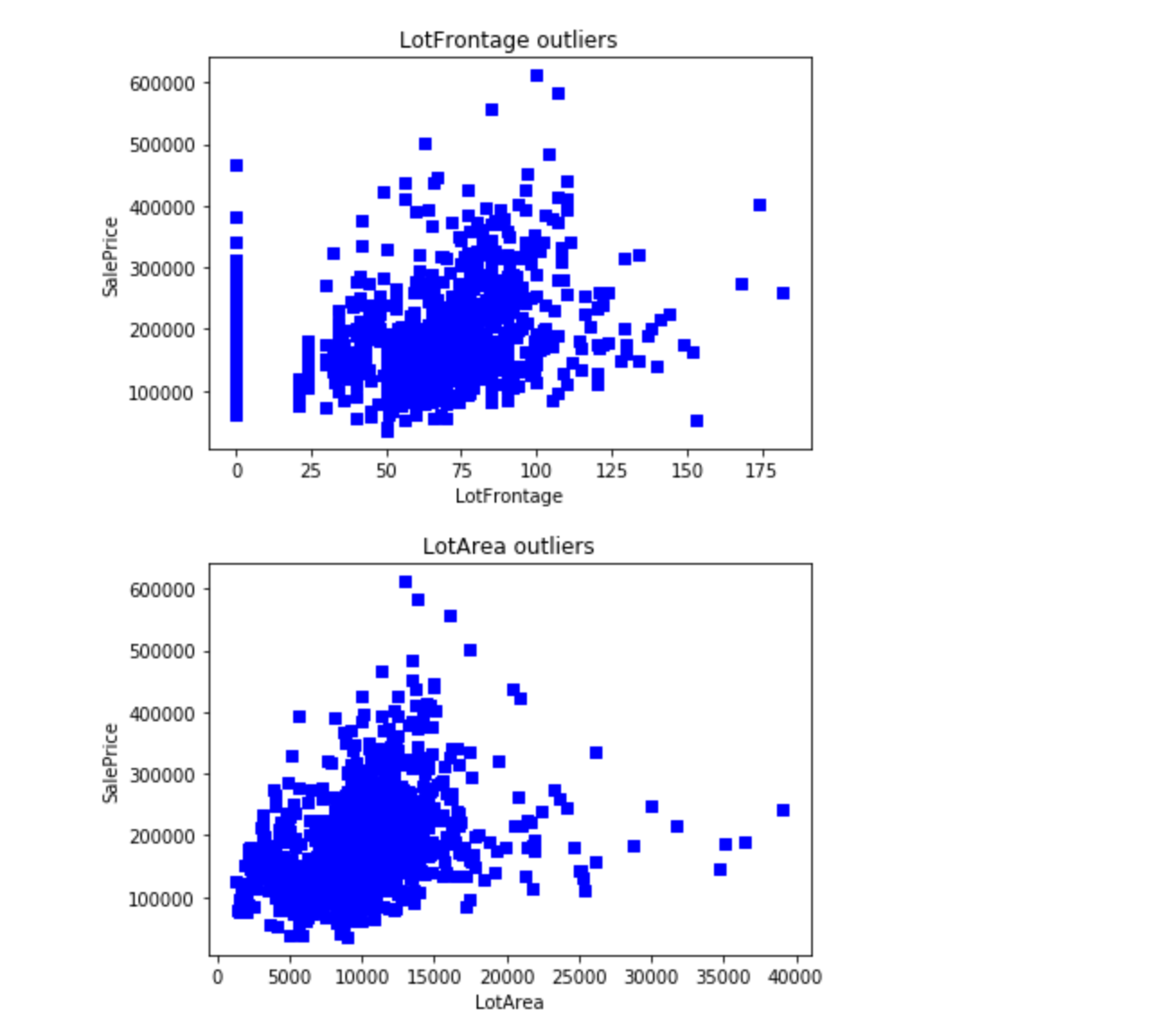
Examples of before and after performing outlier removal:

BEFORE:



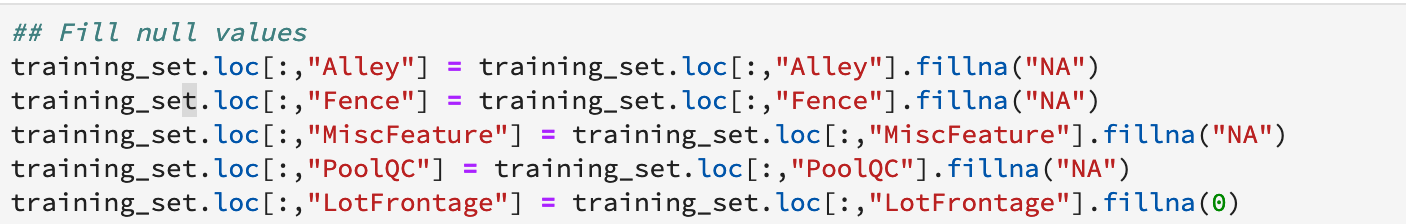


AFTER:



**Dealing with Missing Values:**

Here is a plot of the top features with missing values:



The top 5 features with null values are filled with appropriate null value replacements respectively:

* Numerical null values are assigned 0 if those values are mostly small integers
* Categorical values can be filled with a new category called NA or MISSING

**Machine Learning Models:**

In this project I have used the following ML models

* LASSO regression
* Random Forest

I have tried testing all of the models with various combinations of parameters to determine which model provides the best results in terms of RMSE and Rsquare scores.

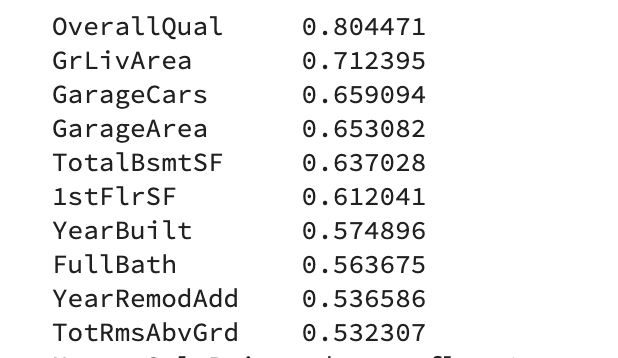
**Feature Selection:**

Feature selection is performed to determine which features have more correlation with the SalePrice and only those are considered for the purpose of fitting the model.

The top 10 features that are closely impacting the SalePrice of a House are taken as the selected features.

The list of top 10 features is given by calculating the correlation coefficient of the features with SalePrice in the train data.

Here is a the top 10 list of features that were produced by the corr() function:



These features become the selected features for fitting into the models.

**Results:**

Among all the Models tested:

* Ridge model produced pretty good results after outlier removal and feature selection with and RMSE score of 0.2066 for the training dataset. But when compared to Random Forest the results of RSME scores it was 3 times more.
* Random Forest produced the second-best results with a RMSE score of 0.07968 for the training dataset.

**Here are the Kaggle score results:**

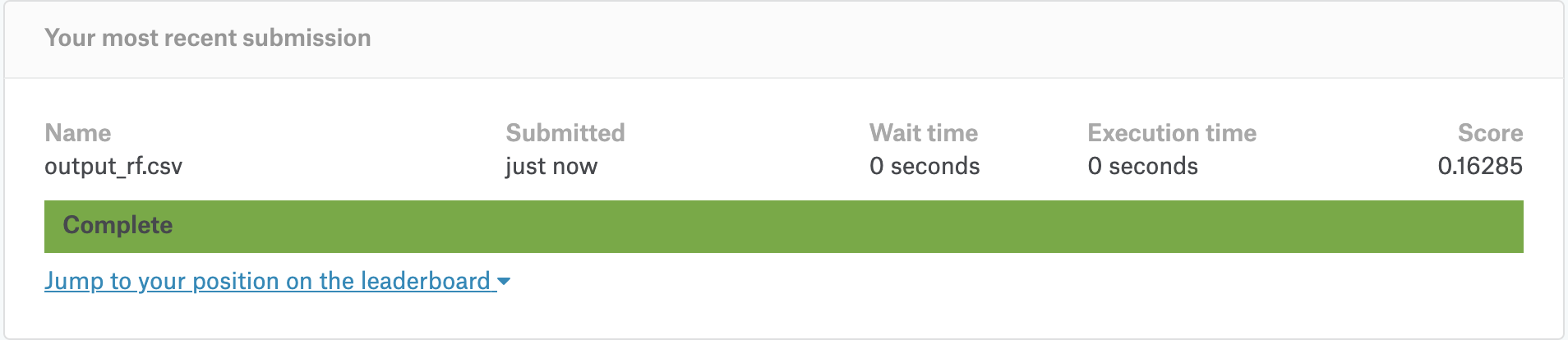


Fig1: Random Forest results

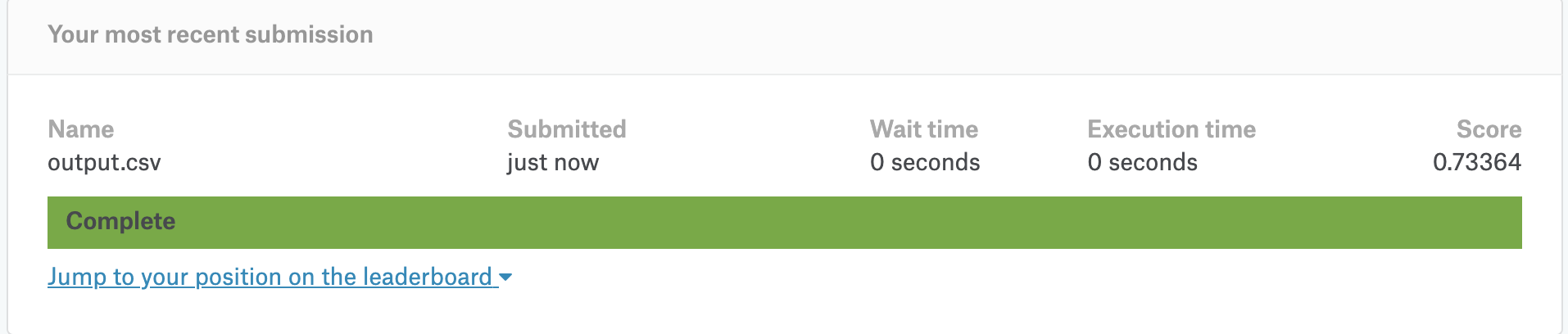


Fig2: Ridge Regression results