

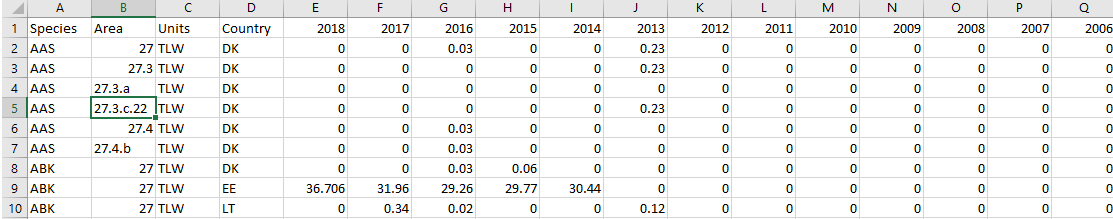
CEBD 1151 - Big Data Analytics for Business

[Team] Assignment – Improve your Model

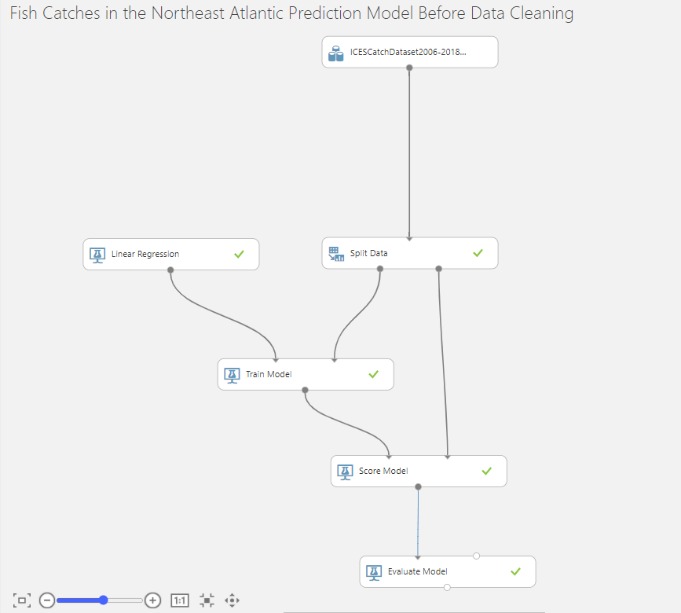
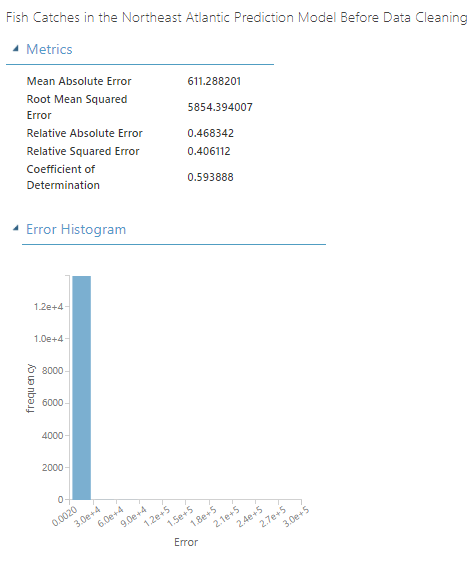
Team members: Michal Velhartický and Omar Bennani

Submission date: 22/11/2020

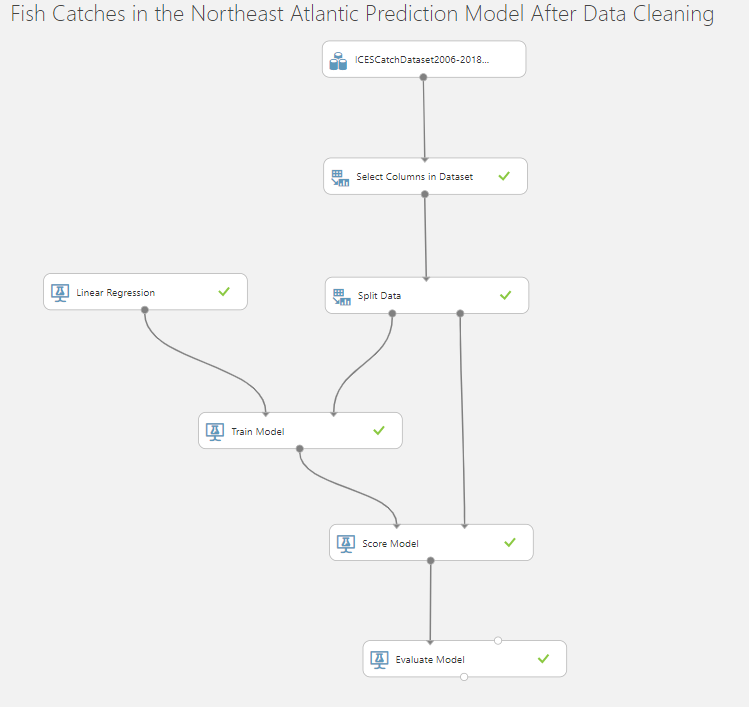
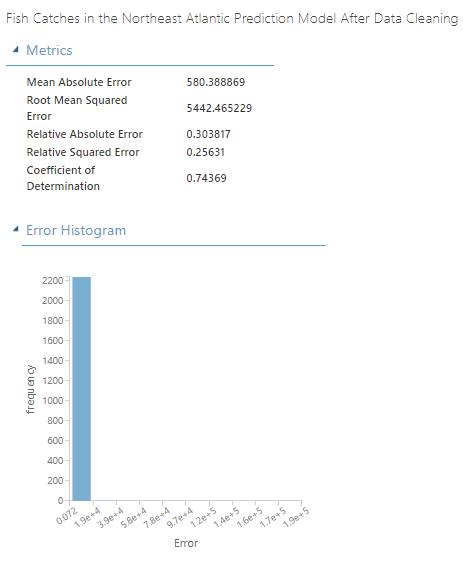
**Data Cleaning**



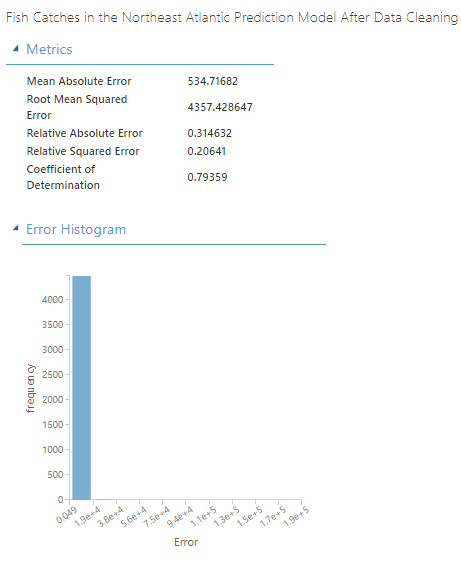
To start the data cleaning, the duplicate data was removed. Part of the information in the dataset is duplicated, because fish catches are counted for each area, sub-area, division, sub-division and unit. The area data includes all the sub-areas and the sub-areas include all the divisions and the divisions include all the sub-divisions and the sub-divisions include all the units. Each row may contain information concerning the area as a whole or only a sub-area, etc. Therefore, we decided only to keep the information classified by sub-area to avoid duplicate information. Another, data cleaning step was to drop the 2018 column, because it contains the letter c in some rows which is not a numerical. Also, the unit column was dropped because it only contains the unit of measurement of the fish catches which is the same for all rows. There are some countries such as Isle of Man, China and Japan which do not have enough data to have an influence, therefore they were removed. The dataset also contains many rows that are all zero values. We also removed these rows, because we do not expect that for a row with only zero values there will be a prediction other than zero.

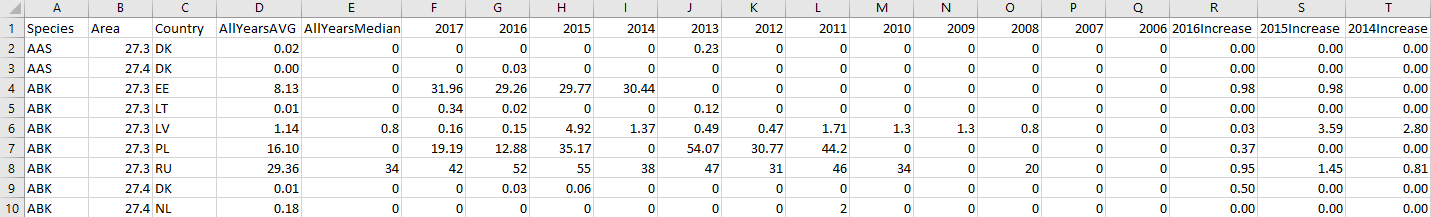
Before data cleaning, by using the original dataset and training a linear model with the target as the year 2017 and a split ratio of training set and validation set of 0.75, the MAE is 611.28, the RMSE is 5854.39 and the R2 is 0.5939. However, after data cleaning and using a linear model with the target as the year 2017, the MAE is 580.39, the RMSE is 5442.46 and the R2 is 0.7437 which is a noticeable improvement from before.

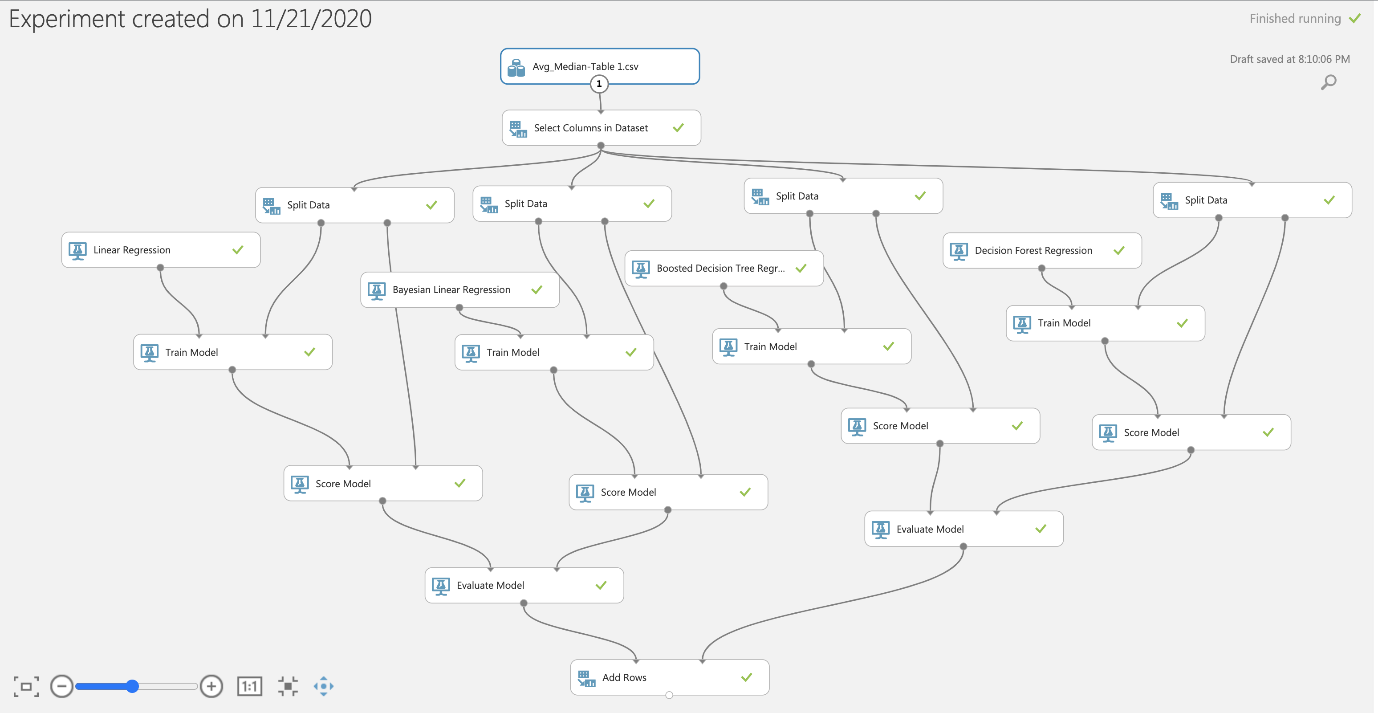
These results were obtained using a split of 0.75. However, with a split 0.5 performance improves with a MAE is 534.72, a RMSE of 4357.42 and the R2 is 0.7936.

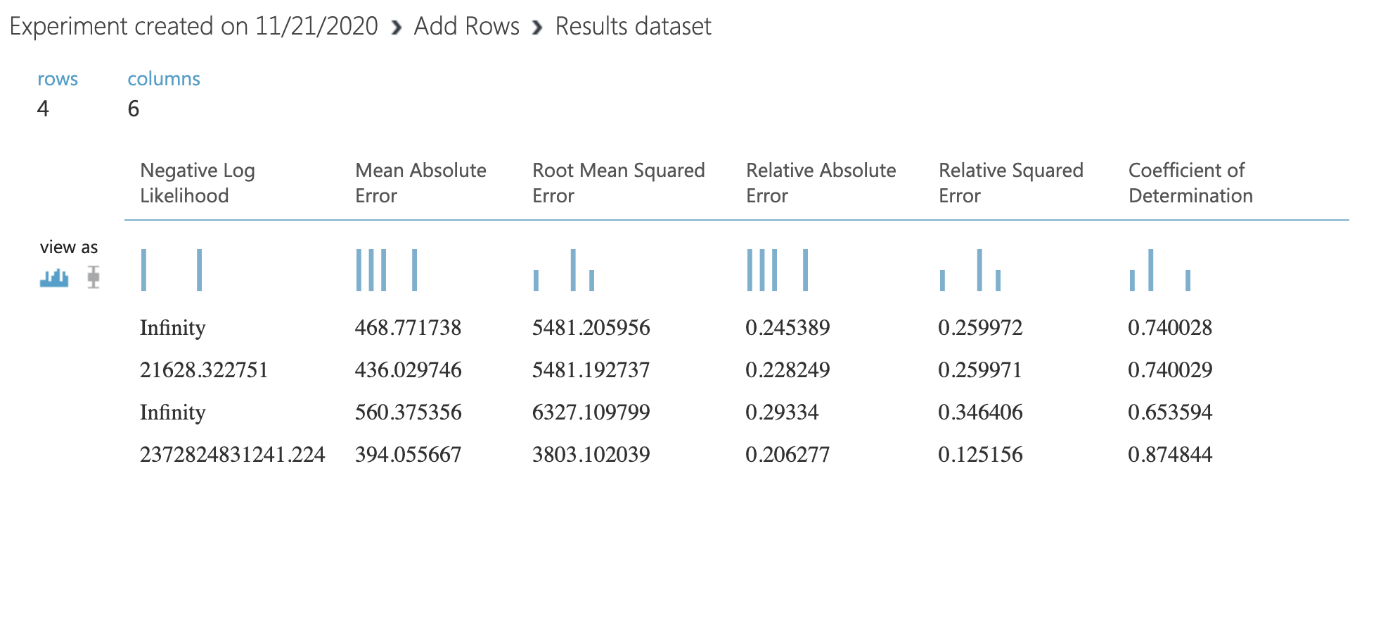


**Feature Engineering and Model Selection**

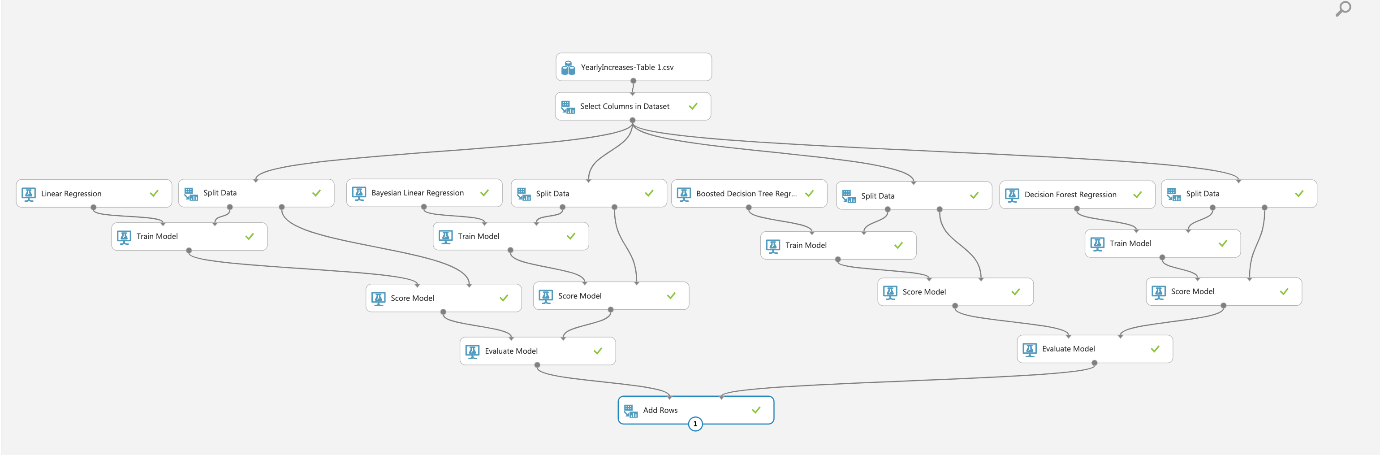


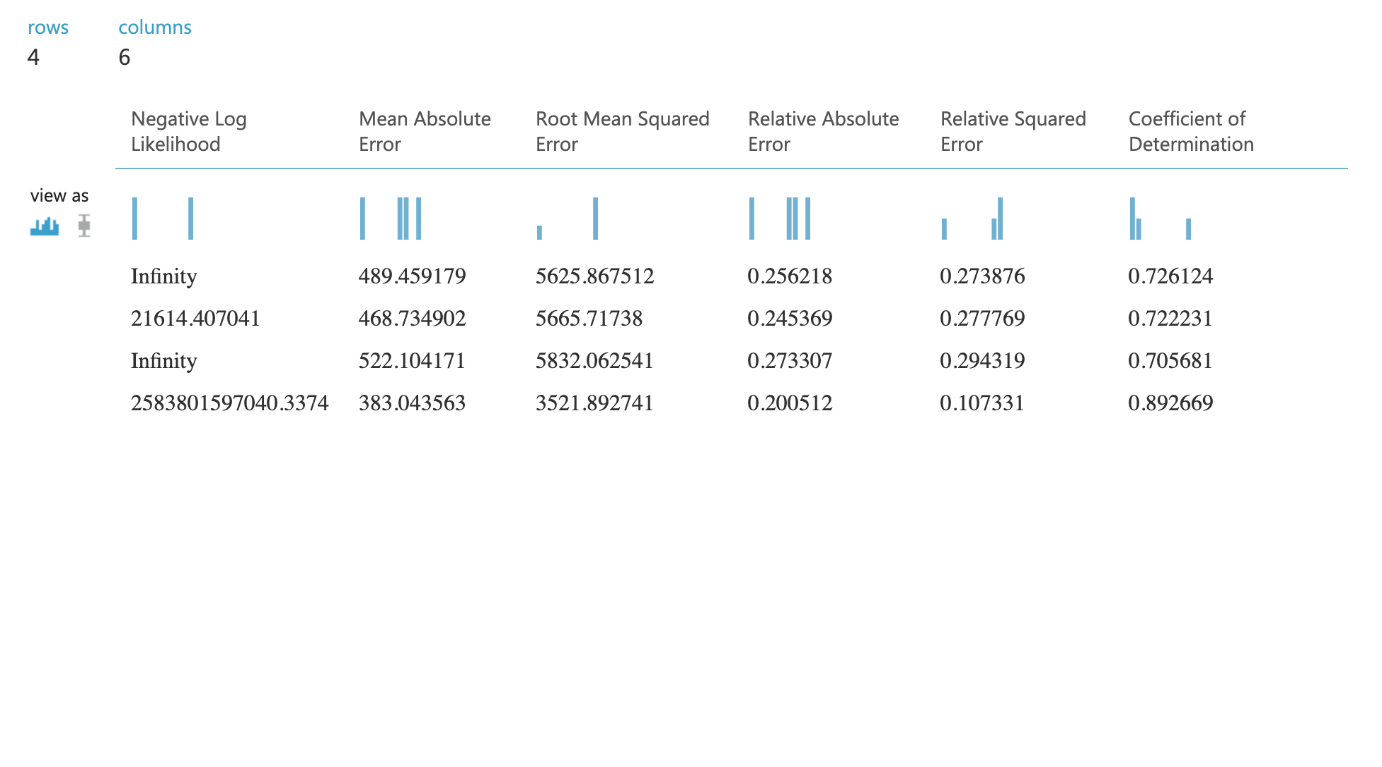
After data cleaning, the process of feature engineering involved adding new features calculated from the original features. Since we are working with a time series data, a useful feature to add is the ratio between the most recent year and the previous year for each year. The most recent year being 2016 and the previous year being 2015. We do not include the target year 2017, since it would be a data leak, because it is information that would not normally be available at prediction time and would artificially improve the performance, but not in reality. Other features added are the average and the median without including the year 2017. For a linear model, these changes did not improve the performance. However, for a boosted tree model and decision forest model, feature engineering improved the results. Looking at the following results, we notice that the decision forest regression model offers the better performance with with a MAE is 394.06, a RMSE of 3803.10 and the R2 is 0.8748.





Following more feature engineering, we changed the ratio between the most recent year and the previous year to be an increase percentage of the most recent year vs the previous year. Using these features instead, the performance increased for the boosted tree regression and the decision forest regression, but it decreased for the linear regression model and the Bayesian linear regression. After these results, we see that the decision forest regression model offers the better performance with a MAE is 383.04, a RMSE of 3521.89 and the R2 is 0.8927.





**Hyperparameter Tuning**

For the final step of our machine learning pipeline, the hyperparameter tuning was used to obtain the optimal set of parameters for the learning model. We used a random sweep mode, a maximum number of runs of 5, we kept the random seed at 0, we used 2017 as the target and the metric we used for evaluating performance is the coefficient of determination. However, the performance did not improve. In order to get the best performance, the hyperparameters tuning used the entire grid testing for each possible combination which required more processing time. It was also evaluating the regression model performance with the coefficient of determination. The decision forest regression performance was slightly improved with a MAE is 371.04, a RMSE of 3406.98 and the R2 is 0.8995.

