Study the scalability and effectiveness of different regression methods in the domain of traffic management

Msc in Data Science — Applied Data Science 2020 - 2021 — 1st Semester Project

Lertas Giorgos
2022202004010

Kratimenos Efstathios
2022202004008

Pitaouli Eftychia
2022202004019

Vakouftsis Athanasios
2022202004002

***Abstract*—This paper is written as part of the Applied Data Science subject's project of the Msc in Data Science program of the University of Peloponnese in collaboration with the National Center of Scientific Research "Demokritos". Since the data available for analysis are constantly growing the need for adapting the current prediction algorithms to perform in reasonable time and with highly accurate results is becoming more and more imperative. The science field of Artificial Intelligence and Machine Learning are presented with the opportunity to achieve great things to improve the human race's lifestyle. The focus point of our study is to study and compare the effectiveness of a variation of regression models in regards to continuously scaling data sets in the domain of airline traffic. After experimenting, the regression methods are compared according to their executing time and a list of algorithm comparing tests.**

***Index Terms*—regression methods, scaling data, algorithm comparison, prediction from data**

I. Introduction

Regression analytics has been the standard approach to modeling the relationship between input and output variables, while recently due to the larger size of the data we aim to incorporate advanced regression analytics capabilities to manage the same predicting results without sacrificing performance. Airline industry is extremely profitable all around the world while the flight data being complicated, and in virtue of this complexity departure and arrival delay are unpredictable. Every year a number of flights get delayed or cancelled due to several reasons. These reasons include weather conditions, security, carrier delays and so on. Reboarding of aircraft due to security breach, defective screening equipments or long lines at screening areas can account for delay in flights. Extreme weather calamities like blizzards, hurricanes and tornados will inevitably lead to flight delays and even cancellations. These parameters are transformed into variables in the context of a regression model which means extra complexity in addition to the large size of the data.

Based on the existing analysis, more productive and competent air traffic management approaches could be planned and implemented. After studying related literature we reach the conclusion that few researches has been done on flight delay forecasting specifically. Some worth mentioning works on flight delay forecasting are: (i) "Flight turnaround time analysis and delay prediction based on bayesian network" by Cao and Lin (ii) "Estimating flight departure delay distributions, A statistical approach with long-term trend and short-term pattern" by Y. Tu, M. O. Ball, and W. S. Jank, [7] where they studied patterns in air traffic delays using statistical approaches. (iii) "Ria-based visualization platform of flight delay intelligent prediction" by R. Yao, W. Jiandong, and D. Jianli, [8] where a RIA (Rich Internet Application)-based visualization platform was created and presented as a solution for flight delay intelligent prediction (iv) "A deep learning approach to flight delay prediction" [9] by J. Kim, S. Choi, S. Briceno, and D. Mavris, where a Recurrent Neural Networks (RNN) where used in addition to a Gradient Boosted Decision Tree approach has been incorporated to predict delays in passenger flights.

II. Data Preparation

A. General Data Information

After exploring the chosen data set, the limited attributes' set decided to used was consisted of the following: 'Day of Week','Airline','Origin Airport','Destination Airport', 'Departure time','Departure Delay','Arrival Time','Arrival Delay'. The scaling of data was made concerning the number of entries rather than choosing the additional complexity of more attributes. The final set consisted of almost 6 million entries where the approach followed was to start with a much smaller subset and measure the performance of the algorithms while increasingly scaling the data. Some additional information about the data are also given such as the fact that the unique origin and destination airports the data are about are 930 in total. Final step of this procedure is a visualization of the "Airlines" attribute relatively to the total count of entries.

B. Data Transformation

Various preprocessing techniques are used in this step for *Data standardization* such as the fit_transform() method of

```
RangeIndex: 5819079 entries, 0 to 5819078
Data columns (total 8 columns):
 #   Column               Dtype
---  ------               -----
 0   DAY_OF_WEEK          int64
 1   AIRLINE              object
 2   ORIGIN_AIRPORT       object
 3   DESTINATION_AIRPORT  object
 4   DEPARTURE_TIME       float64
 5   DEPARTURE_DELAY      float64
 6   ARRIVAL_TIME         float64
 7   ARRIVAL_DELAY        float64
dtypes: float64(4), int64(1), object(3)
memory usage: 355.2+ MB
```

Fig. 1. Final dataset used - Information

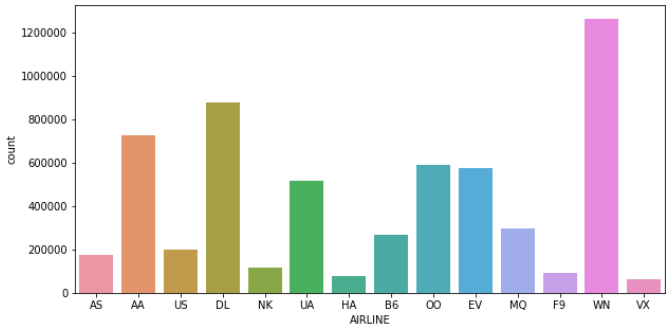

Fig. 2. Count plot with the attribute "Airlines"

scikit-learn library and other functions for scaling features and type transformation. Another necessary step is the resolve of null values to avoid affecting the performance.
Another plot is used to visualize and improve the understanding of the values of the attributes such as the one shown in *figure 3*.

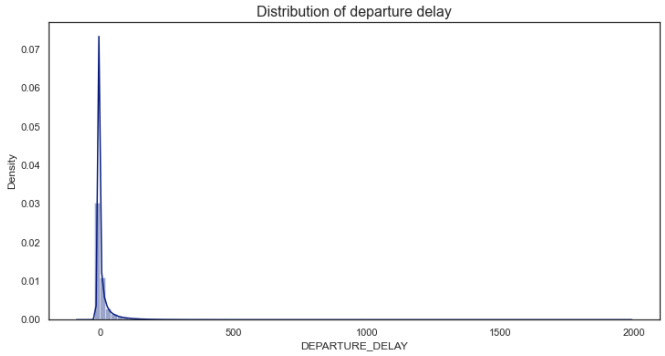

Fig. 3. Distribution of Departure Delay

III. Helper Functions

A. Best learning rate for each scale percentage of the data set

Before moving forward with the main experiment, some helper functions *(Section 4 of the executable)* are created to help automate some standard procedures such as *learn_rate()* used for the decision of the best learning rate used for the *SGDRegressor model* **(Stochastic Gradient Descent)** as well as the *SGD_scaling()* function which makes use of the 1st method to return the best learning rate for each scaled data set. For a list of percentages of the data set and for the chosen learning rates of [0.001, 0.01, 0.1] we have a total of 9 plots presented.

B. Helper Functions for scaling and evaluation

In *Section 5 of the executable*, the useful methods of *scale()* and *evaluation()* are created that are later used for scaling the data by dividing the total of the data set to a subset based on a given percentage to the function as parameter and comparing the algorithms based on a predefined evaluation method-test. The methods effectiveness is compared based on the run time of the algorithm and a selection of statistical tests.

IV. Statistical tests for comparing algorithms

The comparison of samples is made with the use of a number of statistical tests which are presented below:

- *Selection of alpha*
  The p-value can be interpreted in the context of a chosen significance level called alpha. A common value for alpha is 5% or 0.05. If the p-value is below the significance level, then the test says there is enough evidence to reject the null hypothesis and that the samples were likely drawn from populations with differing distributions.
  $p \leq alpha$: reject H0, different distribution
  $p > alpha$: fail to reject H0, same distribution.

- *Friedman test for repeated measurements*
  The Friedman test is a non-parametric statistical test. Similar to the parametric repeated measures ANOVA, it is used to detect differences in treatments across multiple test attempts. The Friedman test tests the null hypothesis that repeated measurements of the same individuals have the same distribution. It is often used to test for consistency among measurements obtained in different ways. For example, if two measurement techniques are used on the same set of individuals, the Friedman test can be used to determine if the two measurement techniques are consistent.
  It ranks the algorithms for each data set separately, the best performing algorithm getting the rank of 1, the second best rank 2.

- *Kruskal-Wallis H-test test*
  The Kruskal-Wallis H-test tests the null hypothesis that

the population median of all of the groups are equal. It is a non-parametric version of ANOVA. The test works on 2 or more independent samples, which may have different sizes. Note that rejecting the null hypothesis does not indicate which of the groups differs. Post hoc comparisons between groups are required to determine which groups are different.

- *Analysis of variance **(ANOVA)***
  Analysis of variance (ANOVA) is a collection of statistical models and their associated estimation procedures (such as the "variation" among and between groups) used to analyze the differences among means. ANOVA is based on the law of total variance, where the observed variance in a particular variable is partitioned into components attributable to different sources of variation. In its simplest form, ANOVA provides a statistical test of whether two or more population means are equal, and therefore generalizes the t-test beyond two means. The one-way ANOVA tests the null hypothesis that two or more groups have the same population mean. The test is applied to samples from two or more groups, possibly with differing sizes.

- *Mann-Whitney U Test*
  The Mann-Whitney U test is a nonparametric statistical significance test for determining whether two independent samples were drawn from a population with the same distribution.The default assumption or null hypothesis is that there is no difference between the distributions of the data samples. Rejection of this hypothesis suggests that there is likely some difference between the samples. More specifically, the test determines whether it is equally likely that any randomly selected observation from one sample will be greater or less than a sample in the other distribution. If violated, it suggests differing distributions.

- *Cohen's d*
  Cohen's d is an effect size used to indicate the standardised difference between two means. It can be used, for example, to accompany reporting of t-test and ANOVA results. It is also widely used in meta-analysis. Cohen's d is an appropriate effect size for the comparison between two means.

- *Wilcoxon rank-sum statistic*
  The Wilcoxon rank-sum test tests the null hypothesis that two sets of measurements are drawn from the same distribution. The alternative hypothesis is that values in one sample are more likely to be larger than the values in the other sample. This test should be used to compare two samples from continuous distributions.

- *Wilcoxon signed-rank test*
  The Wilcoxon signed-rank test is a non-parametric statistical hypothesis test used to compare two related samples, matched samples, or repeated measurements on a single sample to assess whether their population mean ranks differ. The Wilcoxon signed-rank test tests the null hypothesis that two related paired samples come from the same distribution. In particular, it tests whether the distribution of the differences x - y is symmetric about zero. It is a non-parametric version of the paired T-test

- *Independent two-sample t-test*
  This is a two-sided test for the null hypothesis that 2 independent samples have identical average (expected) values. This test assumes that the populations have identical variances by default.

V. Main Part of the Experiment

A. Creating Regression Methods Array

The comparison is made for the following methods *(Section 7 of the executable)*: (i) *Linear Regression*, (ii) *Ridge Regression*, (iii) *XGBoost*, (iv) *Gradient Boost*, (v) *Stochastic Gradient Descent* and each method is executed for the 0.1%, 1.0% and 10.0% of the data.

B. Models comparison with respect to R2 and RMSE

- *R2*
  Let us take a naive approach by taking an average of all the points by thinking of a horizontal line through them. Then we can calculate the MSE *(Mean Square Error)* for this simple model.
  R2 score answers the question that if this simple model has a larger error than the linear regression model. However, it terms of metrics the answer we need is how much larger. The R2 score answers this question. R2 score is 1 — (Error from Linear Regression Model/Simple average model).

- *R-Squared*
  R-Squared is used for evaluating predictions on departure delay of the flight which is a regression machine learning problem.
  R-Squared metric provides an indication of the goodness of fit of a set of predictions to the actual values. The value of R Squared will be between 0 and 1, 0 being no fit and 1 being perfect fit.

VI. Evaluation Results & Running Times

For each regression method mention on *Section V.A* we proceed in running the experiment for the given percentages of the set **[0.1%, 1%, 10%]** and then evaluating the results with the evaluation metrics in *Section V.B*.
The results for each regression algorithm are presented below in the form of a list. The first value represents the evaluation value depending on the specified score metric and the second value the running time.

A. 0.1% of data & R2 Evaluation

1) **Linear Regression:**
   0.883 (0.016), 1.028 sec

2) **Ridge Regression:**
   0.881 (0.013), 1.034 sec

3) **XGBoost:**
   0.910 (0.015), 1.203 sec

4) **Gradient Boost:**
   0.918 (0.012), 3.664 sec

5) **SGD Regression:**
   0.826 (0.062), 1.047 sec

These results from the statistical tests however reject the hypothesis:
All the regression methods (that we use) have equal R2 score.

Friedman test results are:
$Statistics = 18.560$, $p = 0.001$
These values reject the null hypothesis.

Kruskal-Wallis H-test results are:
$Statistics = 14.282$, $p = 0.006$
These values reject the null hypothesis.

ANOVA results are:
$Statistics = 5.639$, $p = 0.003$
These values reject the null hypothesis.

The Algorithms ranking with respect to effectiveness (boxplots) is the following: 1) **Gradient Boosting:** 0.918 (0.012) 2) **XGBoost:** 0.910 (0.015) 3) **Linear Regression:** 0.883 (0.016) 4) **Ridge Regression:** 0.881 (0.013) 5) **Stochastic Gradient Descent:** 0.826 (0.062)

B. 0.1% of data & Negative RMSE Evaluation

1) **Linear Regression:**
   -11.771 (0.396), 1.029 sec

2) **Ridge Regression:**
   -11.913 (0.583), 1.03 sec

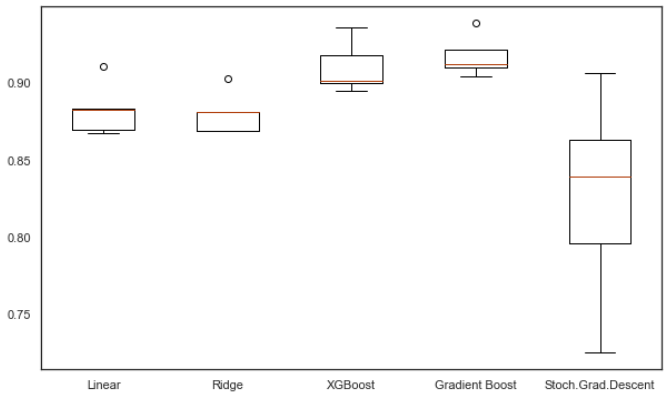

Fig. 4. 0.1% of Data & R2 Score

3) **XGBoost:**
   -10.284 (0.522), 1.237 sec

4) **Gradient Boost:**
   -9.869 (0.456), 3.463 sec

5) **SGD Regression:**
   -13.911 (1.530), 1.041 sec

These results however reject the hypothesis:
Gradient Boosting and Linear Regression have equal RMSE.

The statistical tests are:

Mann-Whitney U test results are:
$Statistics = 0.000$, $p = 0.006$
Wilcoxon rank-sum results are:
$Statistics = -2.611$, $p = 0.009$
Wilcoxon signed-rank results are:
$Statistics = 0.000$, $p = 0.043$
T-test results are:
$Statistics = -6.301$, $p = 0.000$
Another hypothesis is the following:
Stochastic gradient descent and XGBoost regression have equal RMSE.

It is also rejected due to the statistical tests' results presented below.
Mann-Whitney U:
$Statistics = 0.000$, $p = 0.006$
Wilcoxon rank-sum statistic:

$Statistics = 2.611$, $p = 0.009$

Wilcoxon signed-rank test:
$Statistics = 0.000$, $p = 0.043$

T-test:
$Statistics = 4.486$, $p = 0.007$

The Algorithms ranking with respect to effectiveness (boxplots): 1) **Gradient Boosting:** -9.869 (0.456) 2) **XGBoost:** -10.284 (0.522) 3) **Linear Regression:** -11.771 (0.396) 4) **Ridge Regression:** -11.913 (0.583) 5) **Stochastic Gradient Descent:** -13.911 (1.530)

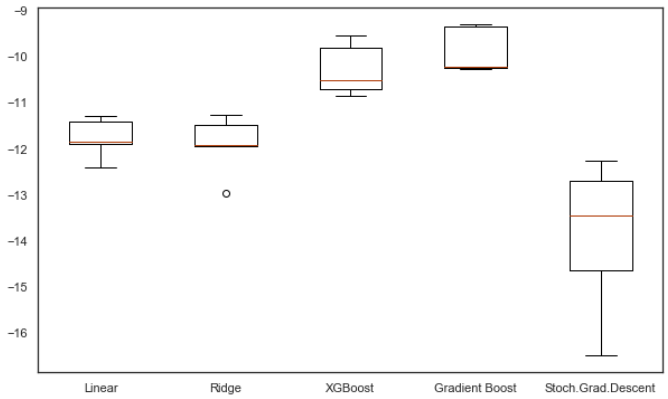

Fig. 5. 0.1% of Data & negative RMSE Score

C. 1% of data & R2 Evaluation

1) **Linear Regression:**
   0.886 (0.003), 1.049 sec

2) **Ridge Regression:**
   0.886 (0.003), 1.045 sec

3) **XGBoost:**
   0.922 (0.004), 2.289 sec

4) **Gradient Boost:**
   0.924 (0.003), 23.143 sec

5) **SGD Regression:**
   0.867 (0.003), 1.185 sec

These results however reject the hypothesis:
All the regression methods (that we use) have equal R2 score. The results from the statistical tests are:

Friedman test:
$Statistics = 19.360$, $p = 0.001$

Kruskal-Wallis H-test:
$Statistics = 21.083$, $p = 0.000$

ANOVA:
$Statistics = 242.727$, $p = 0.000$

The Algorithms ranking with respect to effectiveness (boxplots) is:

1) **Gradient Boosting:** 0.924 (0.003)
2) **XGBoost:** 2.0.922 (0.004)
3) **Linear Regression:** 0.886 (0.003)
4) **Ridge Regression:** 0.886 (0.003)
5) **Stochastic Gradient Descent:** 0.867 (0.003)

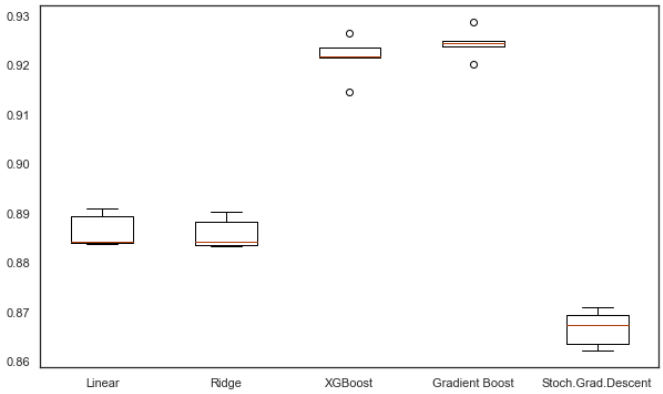

Fig. 6. 1% of Data & R2 Score

D. 1% of data & Negative RMSE Evaluation

1) **Linear Regression:**
   -12.077 (0.055), 1.165 sec

2) **Ridge Regression:**
   -12.107 (0.068), 1.045 sec

3) **XGBoost:**
   -10.036 (0.370), 1.905 sec

4) **Gradient Boost:**
   -9.855 (0.282), 23.498 sec

5) **SGD Regression:**
   -13.093 (0.338), 1.201 sec

These results however reject the following hypothesis:
Gradient Boosting and Linear Regression have equal RMSE.

The statistical tests' results are:
Mann-Whitney U test:
$Statistics = 0.000$, $p = 0.006$

Wilcoxon rank-sum statistic:
$Statistics = -2.611$, $p = 0.009$

Wilcoxon signed-rank test:
$Statistics = 0.000$, $p = 0.043$

T-test:
$Statistics = -15.480$, $p = 0.000$

The following hypothesis is also rejected due to the statistical tests' results.
Stochastic gradient descent and XGBoost regression have equal RMSE.

Mann-Whitney U test:
$Statistics = 0.000$, $p = 0.006$

Wilcoxon rank-sum statistic:
$Statistics = 2.611$, $p = 0.009$

Wilcoxon signed-rank test:
$Statistics = 0.000$, $p = 0.043$

T-test:
$Statistics = 12.186$, $p = 0.000$

The Algorithms ranking with respect to effectiveness (boxplots) is:

1) **Gradient Boosting:** -9.855 (0.282)
2) **XGBoost:** -10.036 (0.370)
3) **Linear Regression:** -12.077 (0.055)
4) **Ridge Regression:** -12.107 (0.068)
5) **Stochastic Gradient Descent:** -13.093 (0.338)

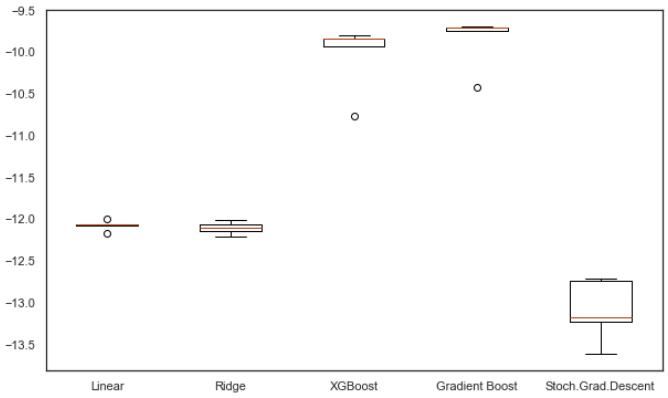

Fig. 7. 1% of Data & negative RMSE Score

E. 10% of data & R2 Evaluation

1) **Linear Regression:**
   0.892 (0.004), 1.565 sec

2) **Ridge Regression:**
   0.892 (0.004), 1.424 sec

3) **XGBoost:**
   0.930 (0.003), 13.264 sec

4) **Gradient Boost:**
   0.931 (0.003), 256.559 sec

5) **SGD Regression:**
   0.869 (0.002), 2.647 sec

However these results reject the following hypothesis:
All the regression methods (that we use) have equal R2 score.

The statistical tests' results are presented below:

Friedman test:
$Statistics = 19.040$, $p = 0.001$

Kruskal-Wallis H-test:
$Statistics = 20.817$, $p = 0.000$

ANOVA:
$Statistics = 278.652$, $p = 0.000$
The Algorithms ranking with respect to effectiveness (boxplots) is:

1) **Gradient Boosting:** 0.931 (0.003)
2) **XGBoost:** 0.930 (0.003)
3) **Linear Regression:** 0.892 (0.004)
4) **Ridge Regression:** 0.892 (0.004)
5) **Stochastic Gradient Descent:** 0.869 (0.002)

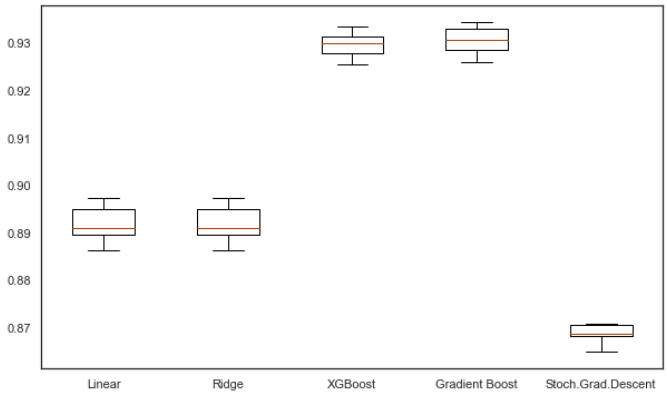

Fig. 8. 10% of Data & R2 Score

F. 10% of data & Negative RMSE Evaluation

1) **Linear Regression:**
   -12.070 (0.069), 1.441 sec

2) **Ridge Regression:**
   -12.071 (0.068), 1.256 sec

3) **XGBoost:**
   -9.733 (0.089), 9.69 sec

4) **Gradient Boost:**
   -9.672 (0.094), 256.049 sec

5) **SGD Regression:**
   -13.306 (0.145), 3.083 sec

However these results reject the following hypothesis:
Gradient Boosting and Linear Regression have equal RMSE.

The Statistical tests' results are:

Mann-Whitney U test:
$Statistics = 0.000$, $p = 0.006$

Wilcoxon rank-sum statistic:
$Statistics = -2.611$, $p = 0.009$

Wilcoxon signed-rank test:
$Statistics = 0.000$, $p = 0.043$

T-test:
$Statistics = -41.312$, $p = 0.000$

Hypothesis **Stochastic gradient descent and XGBoost regression have equal RMSE** is also rejected due to the following results.

Mann-Whitney U test:
$Statistics = 0.000$, $p = 0.006$

Wilcoxon rank-sum statistic:
$Statistics = 2.611$, $p = 0.009$

Wilcoxon signed-rank test:
$Statistics = 0.000$, $p = 0.043$

T-test:
$Statistics = 41.919$, $p = 0.000$
The Algorithms ranking with respect to effectiveness (boxplots) is:

1) **Gradient Boosting:** -9.672 (0.094)
2) **XGBoost:** -9.733 (0.089)
3) **Linear Regression:** -12.070 (0.069)
4) **Ridge Regression:** -12.071 (0.068)
5) **Stochastic Gradient Descent:** -13.306 (0.145)

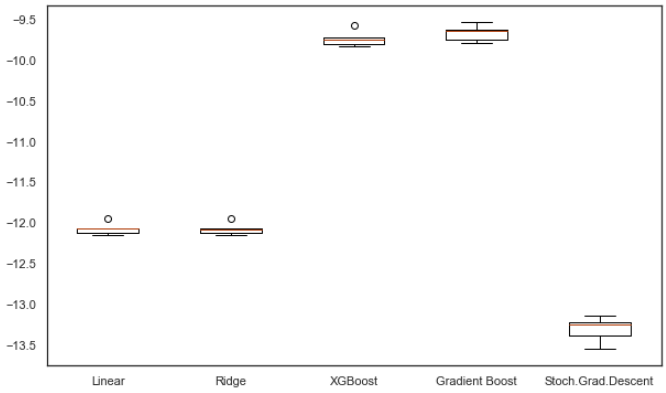

Fig. 9. 10% of Data & negative RMSE Score

VII. Hypothesis' Analysis

A. Null Hypothesis 1

The first null hypothesis is: **s we increase the instances of the data set, then the execution time remains the same**.
It is clear that when we increase the data of the sample, the execution time of the models is increased too and this happens at every model that is used on this research. It is tested for five algorithms twice, so the null hypothesis is rejected.

B. Null Hypothesis 2

The second null hypothesis is: **Gradient Boosting and Linear Regression have equal RMSE(root mean square error)**.
From the comparison of Gradient Boosting and Linear Regression, from the statistic methods the null hypothesis is rejected which have equal RMSE.
The ranking of the algorithms is:

1) **Gradient Boosting**
2) **XGBoost**
3) **Linear Regression**
4) **Ridge Regression**
5) **Stochastic Gradient Descent**

So the null hypothesis is rejected, because Gradient Boosting has better results that Linear Regression.

C. Null Hypothesis 3

The third null hypothesis is: **All the regression methods (that we use) have equal R2 score**. The comparison of the algorithms shown that XGBoost Regression was the obvious choice with better R Square value and negative RMSE in all scales. In this Null Hypothesis there were rejections with Friedman test, Kruskal-Wallis, H-test and ANOVA. Also, in boxplots the algorithms have not the same scores as a result the alorithms don't have equal R2 score. So the null hypothesis is rejected.

D. Null Hypothesis 4

The fourth null hypothesis is: **Stochastic Gradient Descent and XGBoost regression have equal RMSE**.
In this null hypothesis it is clear that Mann-Whitney U test, Cohen's d, Wilcoxon rank-sum statistic, Wilcoxon signed-rank test, T-test, all of them reject null hypothesis, so this results that every algorithm don't have equal RMSE. Also, doing the comparison of the algorithms with RMSE, noted that the methods had the same ranking with R2 score. The ranking of the algorithms is:

1) **Gradient Boosting**
2) **XGBoost**
3) **Linear Regression**
4) **Ridge Regression**
5) **Stochastic Gradient Descent**

So the null hypothesis is rejected with Stochastic Gradient Descent and XGBoost Regression having different RMSE. XGBoost Regression has smaller RMSE than Stochastic Gradient Decent and the best compared to other Regression algorithms.

E. Null Hypothesis 5

The fifth null hypothesis is: **When using more instances from the data, then regression methods achieves less R2 score**.

As seen on the table with R2 scores, there is a difference with the increase of the data in sampling, which is that the R2 score also increase. Observing the results ranking of the algorithms have the same with RMSE:

1) **Gradient Boosting**
2) **XGBoost**
3) **Linear Regression**
4) **Ridge Regression**
5) **Stochastic Gradient Descent**

So the null hypothesis is rejected because the R2 score increase with more data.

CONCLUSIONS

To sum up, it is evident by taking into account the collective results that the best overall algorithm for efficiently computing the delay (in airlines traffic specifically for our case) is Gradient Booster since it achieves the best results in most cases in regards to the evaluation score as well as the scale of the data. However, further studies are necessary and a more robust prediction algorithm is yet to be developed.

REFERENCES

[1] "Scalability of Machine Learning Algorithms", Georgios Paliouras Department of Computer Science, 1993

[2] "A Regression-Based Approach to Scalability Prediction", Bradley J. Barnes, Barry Rountree, David K. Lowenthal, Jaxk Reeves, Bronis de Supinski, Martin Schulz

[3] "Large-scale predictive modeling and analytics through regression queries in datamanagement systems", Christos Anagnostopoulos, Peter Triantafillou, 2018

[4] "A Statistical Approach to Predict Flight Delay Using Gradient Boosted Decision Tree", Suvojit Manna, Sanket Biswas, Riyanka Kundu Somnath Rakshit, Priti Gupta, Subhas Barman, 2017 International Conference on Computational Intelligence in Data Science(ICCIDS)

[5] "ETCPS: An Effective and Scalable Traffic Condition Prediction System", Dong Wang, Wei Cao, Mengwen Xu, Jian Li

[6] W.-d. Cao and X.-y. Lin, "Flight turnaround time analysis and delay prediction based on bayesian network," Computer Engineering and Design, vol. 5, pp. 1770–1772, 2011.

[7] "Estimating flight departure delay distributionsa statistical approach with long-term trend and short-term pattern", Y. Tu, M. O. Ball, and W. S. Jank, Journal of the American Statistical Association, vol. 103, no. 481, pp. 112–125, 2008.

[8] "Ria-based visualization platform of flight delay intelligent prediction", R. Yao, W. Jiandong, and D. Jianli, Computing, Communication, Control, and Management, 2009. CCCM 2009. ISECS International Colloquium on, vol. 2. IEEE, 2009, pp. 94–97.

[9] "A deep learning approach to flight delay prediction", Y. J. Kim, S. Choi, S. Briceno, and D. Mavris, Digital Avionics Systems Conference (DASC), 2016 IEEE/AIAA 35th. IEEE, 2016, pp. 1–6.

[10] Traffic Management System Performance Using Regression Analysis, D. Levinson, W. Chen, 2006