# Prediction and Data Analysis of Horse Racing in Turkey

Çağatay Gülten

cagataygulten@gmail.com

August 2021

**Abstract:** Horse racing is a social demand of modern societies. Besides other sports it is presented by horses and so that it is an eye-catching sport. Gambling part of this sport is also a common way to make it more popular. Success on horse racing gambling requires mostly knowledge of horses. Thus, it is suchlike an experience earned over time by watching the races. But there is also a statistical way to handle it. In this paper several concepts and inferences in horse racing in Turkey are considered by processing statistical data and an algorithm that makes prediction by using regression models is introduced.

## 1. Introduction

Horse racing is a competition involving many horses ridden by jockeys on a specific type of surface and over a distance. The aim of this competition is to find the fastest horse in that situation and condition. But of course, all of the horses do not have the same properties. There are different properties of horses that makes them clearly different such as body power, genders, breeds, ages. Approximate results are expected to make the race more competitive and fair. For this reason, before races, institutions apply some methods to make them perform more equally such as putting additional weights and setting race prerequisites.

When making decisions for a race, all external conditions should be considered as same. Properties such as prerequisites, breeds, surface, distance are always same in a race. Only horses' body power specifies the results. The most common way to decide most powerful horse is considering its previous races and gathering statistics.

## 2. Data Representation

In this project, the data collected by TJK (Turkish Jockey Club) are used. The data consist of all races in Turkey from 1988 until now and has totally more than 1,100,000 entries. Raw data have surface, distance, city, weight, age, breed, gender, horse id, mother id, father id, jockey id, final place and time values. Distance values were turned into categorical values as short (<1450 meters), middle (1450 – 1950 meters) and long (>1950 meters).

There are two horse breeds run in the official races in Turkey. Thoroughbred and Arabian horses, but Turkish horse racing society uses "thoroughbred" word for all kinds of racing horses and calls two breeds as "English" and "Arabian".

In this part, effects of distance, surface, city, additional weights, breed and gender on horses' run time are analysed and correlations are observed.

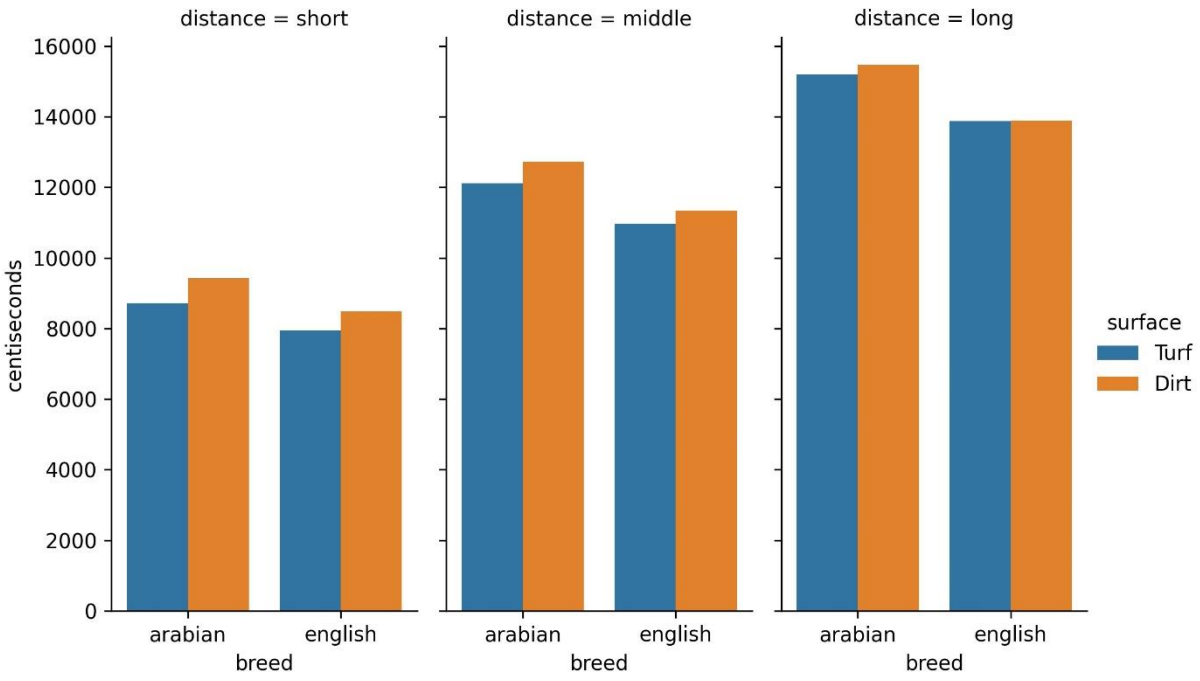


*Figure 1: Average run times of two horse breeds on different distances and surfaces*

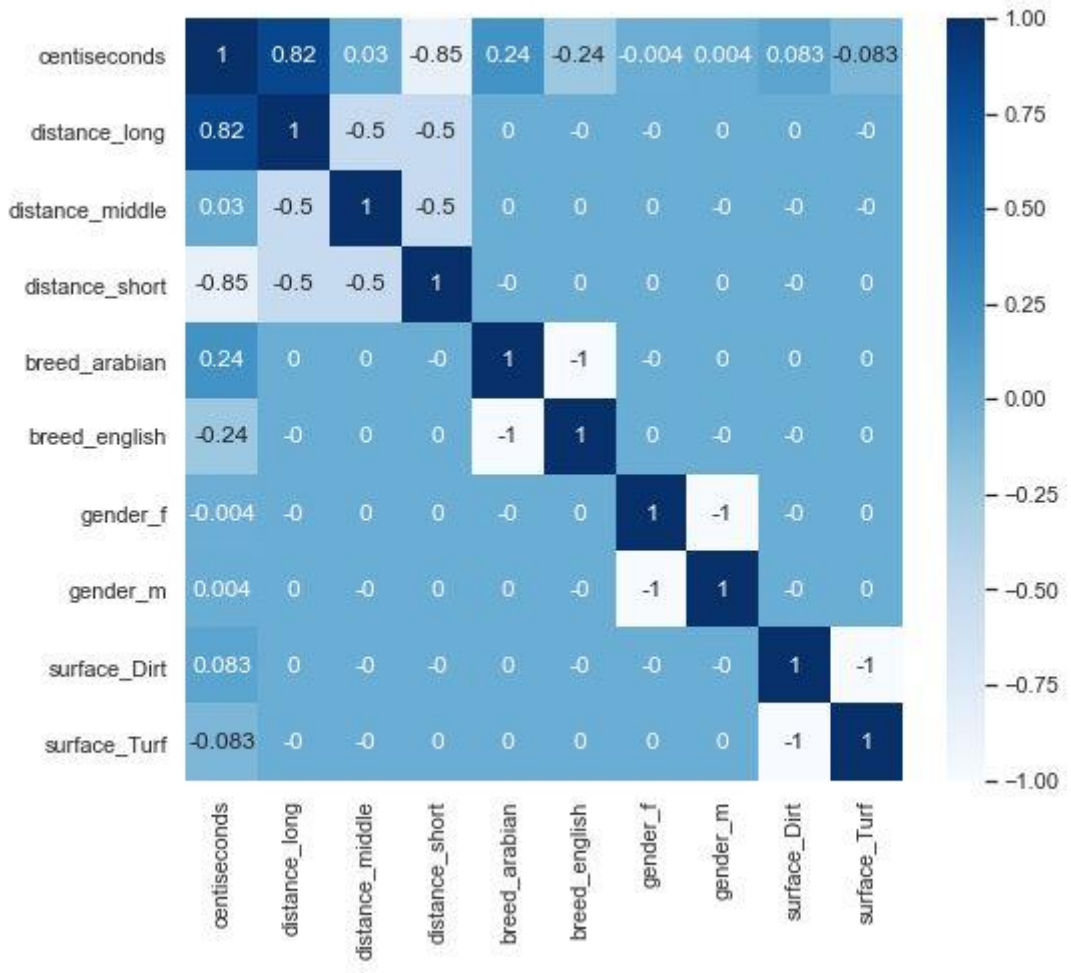


*Figure 2: Correlation heat map*

As expected also in terms of physics, the distance makes the most difference. But the distance has also side effects when it is considered with other properties as a combination. For example, as seen in Figure 1, surface is less effective at long distances than short distances. As correlation matrix in Figure 2, surface has only 8% effect on time. That is caused by friction made on horses' hooves which is greater on dirt surfaces than turf surfaces.

Genders has almost no effect on the time. So it is not possible to say a gender has more advantages and also in Turkey it is allowed to run different genders in a same race, but sometimes there is a gender condition in prerequisites.

When the breeds are considered, it is clear that English breeds (thoroughbred) performs better performance in all conditions. Likely it is caused by genetic factors on thoroughbreds' bodies inherited from origins.

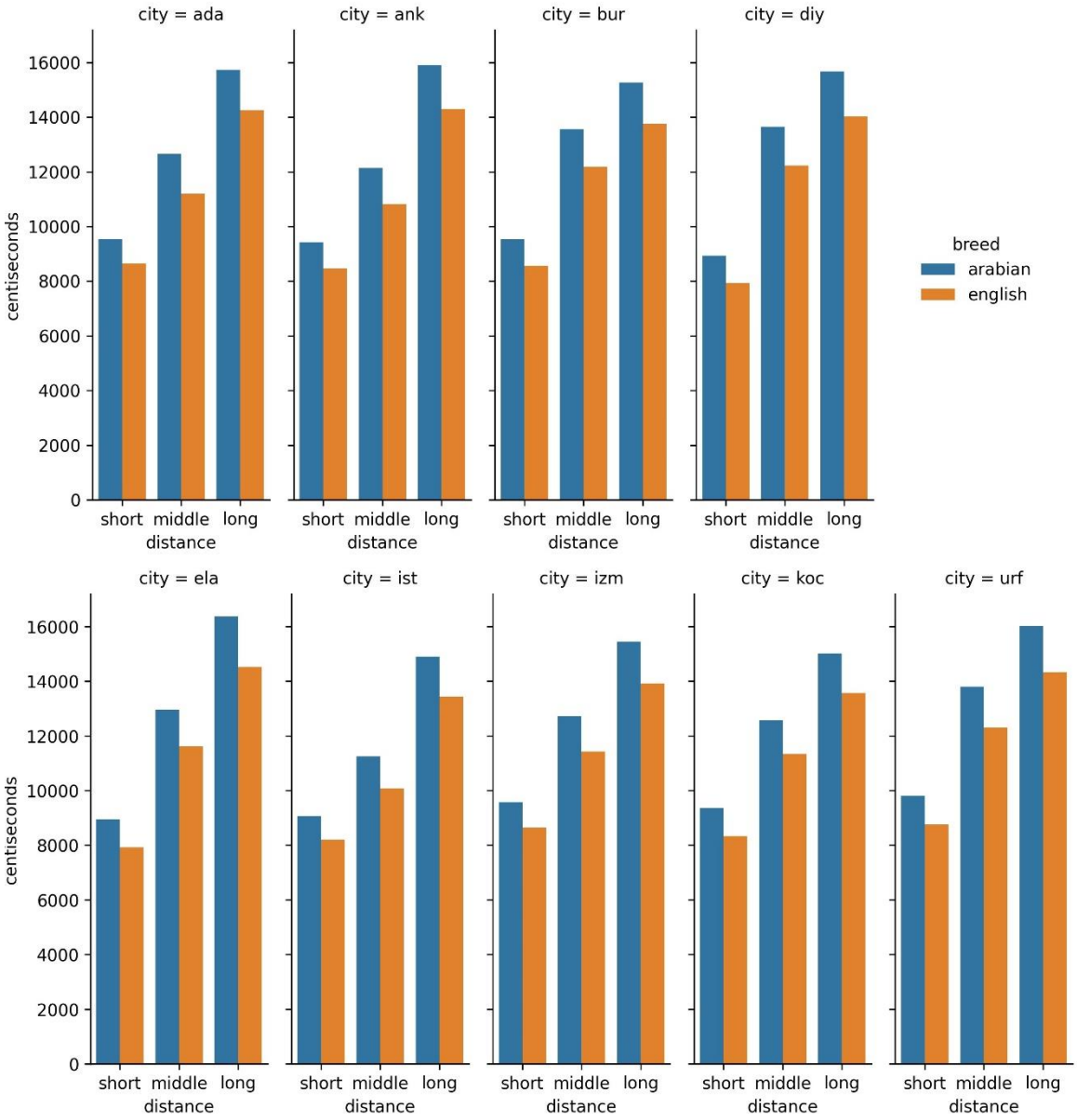


*Figure 3: Average times of races on dirt surface in different cities*

If the differences of cities which have dirt surface field are considered, it can be said that dirt in Elazığ and Urfa is more effective on horses' hooves than other cities and that causes higher average time. Besides that, there is no clear difference between turf surfaces in several cities as it seen in Figure 4.

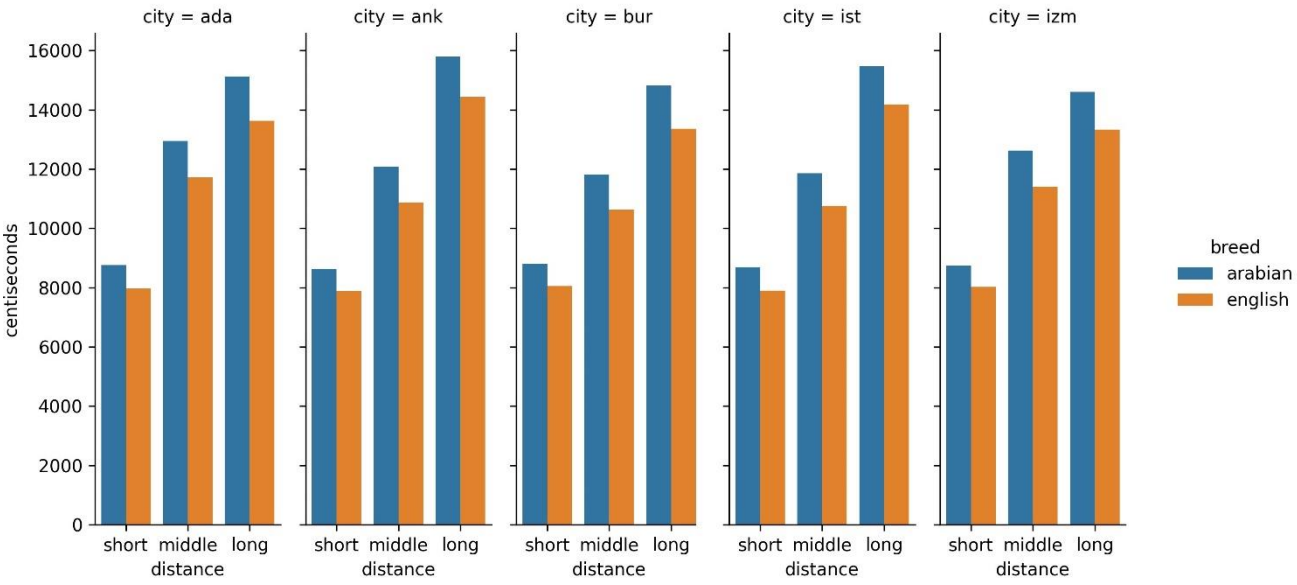


*Figure 4: Average times of races on turf surface in different cities*

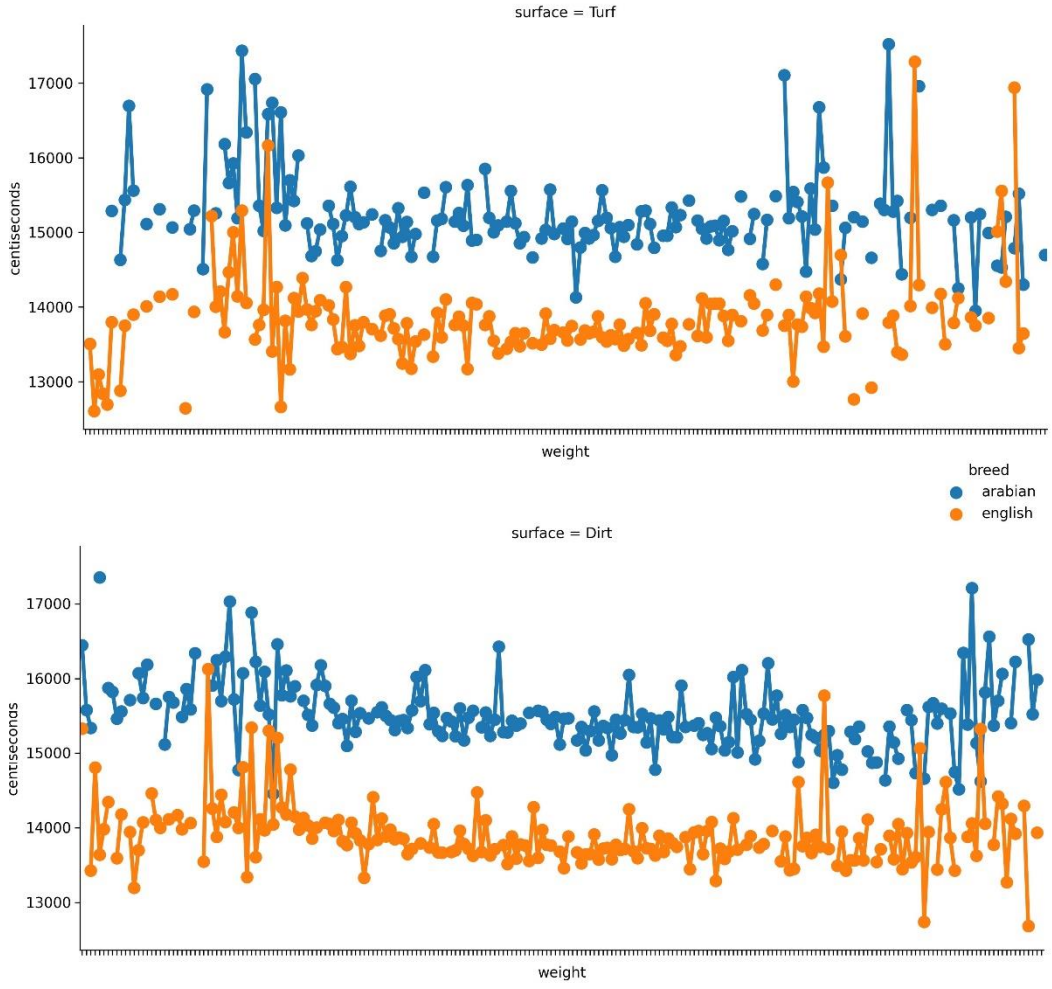


*Figure 5: Average times of long distance races on different surfaces with additional weights*

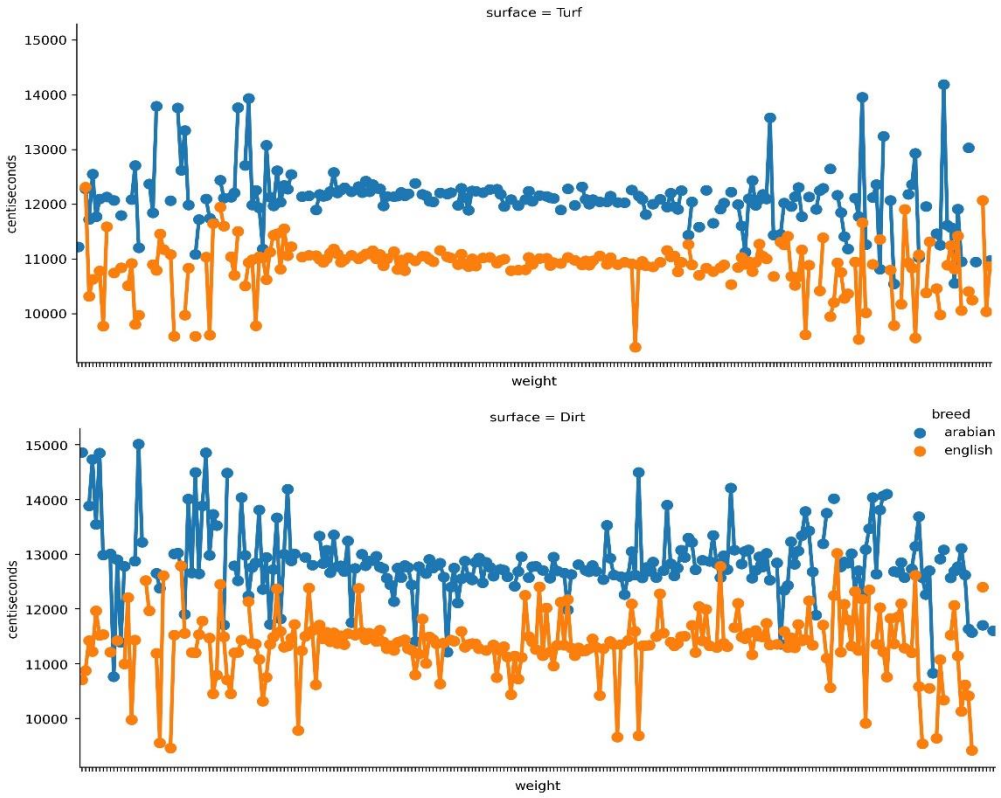


*Figure 6: Average times of middle distance races on different surfaces with additional weights*

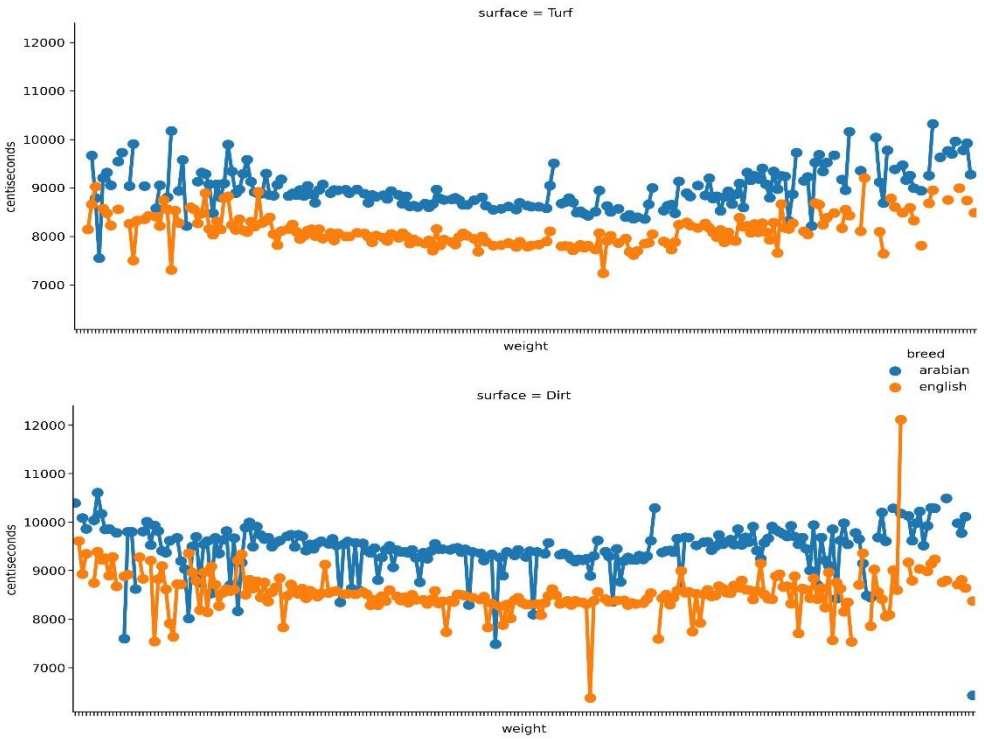


*Figure 7: Average times of short distance races on different surfaces with additional weights*

Lastly, additional weights are placed on horses to equalize horse performances. According to that situation, it is expected to no correlation between time and additional weights. As it is seen in Figure 5, 6 and 7, the mostly used additional weights are in the middle part of diagrams and there is a horizontal linearity. Side parts of diagrams (low and high weights) represent rarely added additional weights and due to lack of data it is appeared irregularly.

## 3. Data Preprocessing for Prediction Algorithm

Data preprocessing is required to obtain statistical data by managing features. All horses' previous races statistics presents their body performances.

By filtering data for each jockey and horse:

1. top 1,2,3 and 4 percentages of jockeys,
2. top 1,2,3 and 4 percentages of horses on the current surface,
3. top 1,2,3 and 4 percentages of horses on the current distance,
4. top 1,2,3 and 4 percentages of horses' mother's all children on the current surface,
5. top 1,2,3 and 4 percentages of horses' mother's all children on the current distance,
6. top 1,2,3 and 4 percentages of horses' mother's all children,
7. top 1,2,3 and 4 percentages of horses' father's all children on the current surface,
8. top 1,2,3 and 4 percentages of horses' father's all children on the current distance,
9. top 1,2,3 and 4 percentages of horses' father's all children

are obtained.

The aim of getting mother and father's children statistics is to consider genetic factors inherited from horses' parents. After that all of these percentages are turned into single weighted average value. "Pct." stands for percentage.

$$Weighted\ average = \frac{4x\ Top\ 1\ Pct. + 3x\ Top\ 2\ Pct. + 2x\ Top\ 3\ Pct. + Top\ 4\ Pct.}{10}$$

Python script of this feature managing process is accessible in the project folder ("horce_racing_feature_manager.py").

Finally, after removing outliers, the data consist of only 9 weighted average values that represents all statistics in the current race conditions.

```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 327148 entries, 0 to 1138976
Data columns (total 9 columns):
 #   Column                    Non-Null Count   Dtype
---  ------                    --------------   -----
 0   jockey_wa                 327148 non-null  float64
 1   thoroughbred_surface_wa   327148 non-null  float64
 2   mother_surface_wa         327148 non-null  float64
 3   father_surface_wa         327148 non-null  float64
 4   thoroughbred_distance_wa  327148 non-null  float64
 5   father_distance_wa        327148 non-null  float64
 6   mother_distance_wa        327148 non-null  float64
 7   mother_general_wa         327148 non-null  float64
 8   father_general_wa         327148 non-null  float64
dtypes: float64(9)
memory usage: 25.0 MB
```

*Figure 8: Train set info of short distances and dirt surfaces*

## 4. Prediction Algorithm Architecture

After getting train sets for all conditions, they were fit into 5 different regression models. They are Gradient Boosting Regression, K-Nearest Neighbours Regression, Lasso Regression, Ridge Regression and a containing all these models a Voting Regressor.

Prediction algorithm scrapes desired race's required data from a website by using selenium library and chrome webdriver, creates a test set and fits to all 5 models. As a result, it plots a table that contains predicted time values for each horse in that race by coloring cells to notice performances easily.

| | GradientBoostReg | KNNReg | LassoReg | RidgeReg | VotingReg |
|---|---|---|---|---|---|
| 1 | 11832.08 | 12043.5 | 11913.34 | 11819.95 | 11902.22 |
| 2 | 10989.85 | 10665.1 | 11598.69 | 11630.24 | 11220.97 |
| 3 | 12057.0 | 12058.1 | 12457.93 | 12463.37 | 12259.1 |
| 4 | 12463.05 | 13006.2 | 12317.44 | 12263.05 | 12512.44 |
| 5 | 11442.52 | 11860.2 | 11686.01 | 11644.38 | 11658.28 |
| 6 | 11508.55 | 11164.4 | 11761.4 | 11560.16 | 11498.63 |
| 7 | 12077.15 | 11816.4 | 11995.2 | 11970.33 | 11964.77 |
| 8 | 11739.1 | 11603.8 | 11957.84 | 12015.86 | 11829.15 |
| 9 | 11949.54 | 12505.3 | 12230.51 | 12243.48 | 12232.21 |

*Figure 9: Example of prediction output table*

## 5. Discussion

The results of predictions do not represent actual expected times of the given race, because as written in data representation part, distance variables are converted to categorical variables to pack races and consider together. Therefore, regression models calculate it as an average time of given distance. Namely, prediction outputs should be only used for comparison of horses' performance. Green cells means better performance.

All scripts, plots and a user manual are available in this project's folder, but the main data are not permitted to be published, thus a demonstration data is created for the users wanting to try prediction algorithm. This demo data only contain races in one week period between 12/04/2021 and 18/04/2021. So the user should select a day in this period. All information is given in the user manual file.

The accuracy of the prediction algorithm is quite enough to give a clue to a person interested in horse racing. It is not expected to get 90% accuracy from an algorithm that makes a prediction for a gamble. All regression models make consistent predictions and sometimes KNN regression model made some particular decisions. The output table contains 5 all predictions of models to analyse decisions entirely. Voting regressor is the most steady model as expected, it is recommended to make decisions regarding the output of voting regressor.

## 6. Conclusion

To conclude, a person who has never watched any horse race can also make good predictions through statistics, a programming language along with machine learning algorithms and math. This project gives a clue based on these instruments for horse racing followers and by analysing horse racing data, makes some inferences to add cultural knowledge in the horse racing community in Turkey.

## References

TJK (Turkey Jockey Club), https://www.tjk.org/