**Second Place Prediction Q2 Documentation**

Question:

(2) [25pts] Now we move on to prediction. Fit a model using data from 1950:2010, and predict drivers that come in second place between 2011 and 2017. [Remember, this is a predictive model where variables are selected as the subset that is best at predicting the target variable and not for theoretical reasons. This means that your model should not overfit and most likely be different from the model in (1).]

From your fitted model:

●  describe your model, and explain how you selected the features that were selected

●  provide statistics that show how good your model is at predicting, and how well it performed predicting second places

in races between 2011 and 2017

●  the most important variable in (1) is bound to also be included in your predictive model. Provide marginal effects or

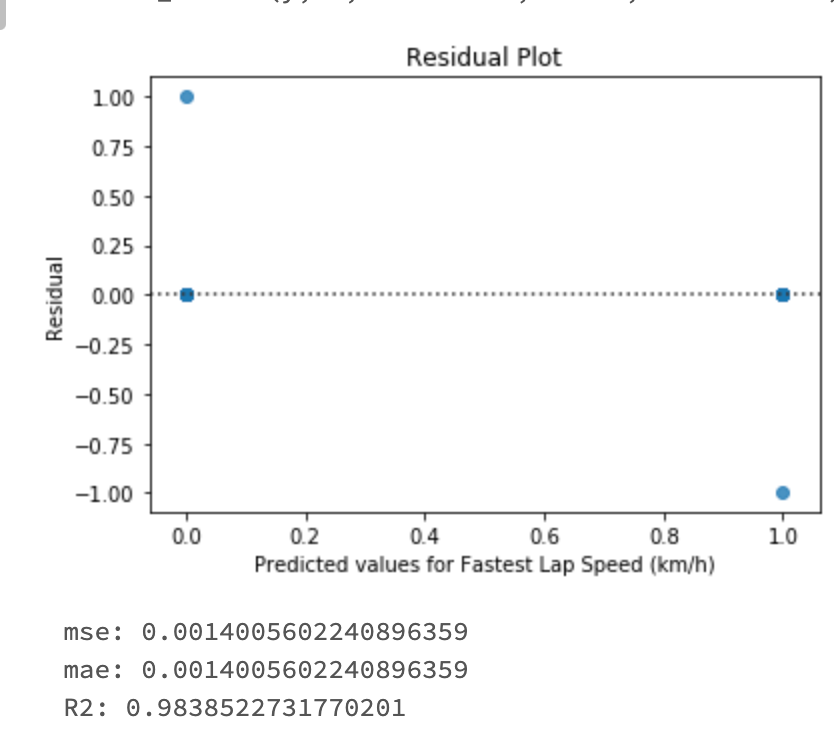
some metric of importance for this variable and make an explicit comparison of this value with the values that you obtained in (1). How different are they? Why are they different?

Data preprocessing:

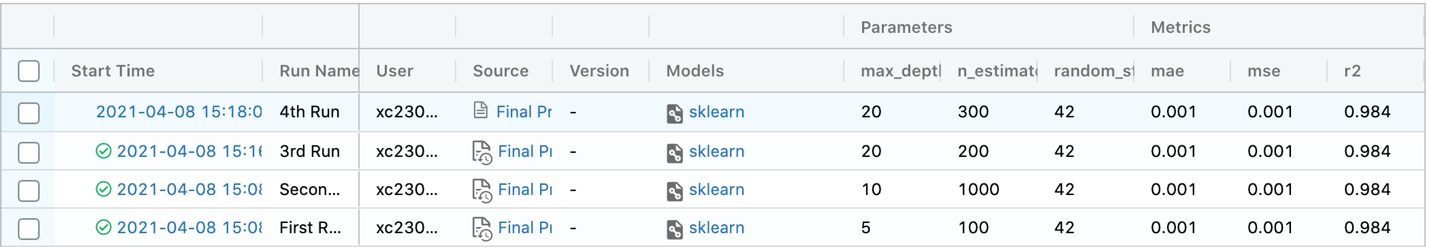
* Use pre-processed dataset: race, driver, constructors and results
* Create binary variable **second\_place** to indicate 2nd place finish
* Create **age\_as\_of\_race** through dob column
* Create binary variables for **constructorRef** with top 5 most wins
* Create binary variables for top 10 **grid** positions
* Turn **milliseconds**, **fastestLap**, **rank and fastestLapSpeed** into numeric datatype

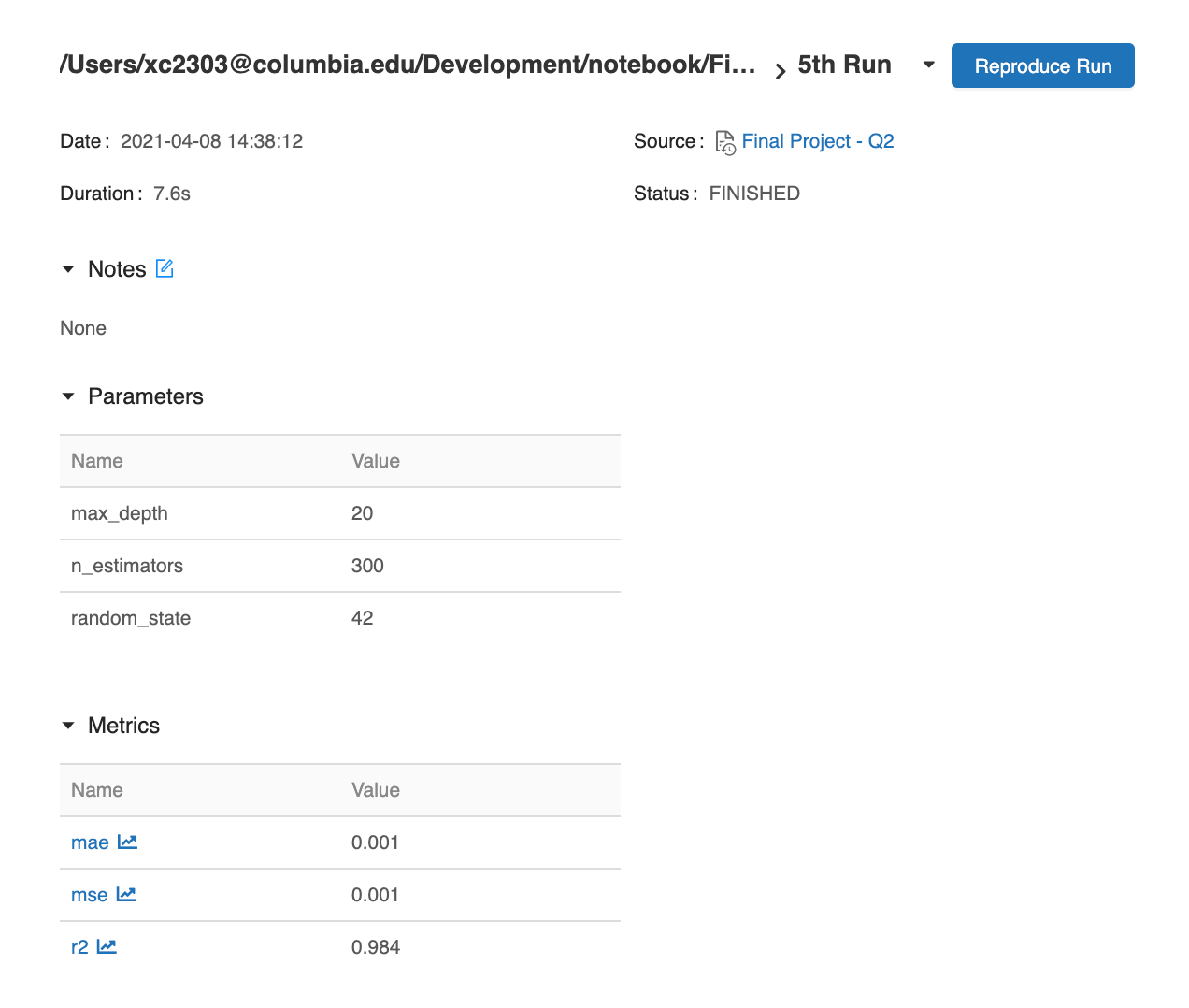
Prediction Process:

* Feature selection: drop non-numeric columns. Feature selected for prediction include: race\_year, points, raceId, milliseconds, second\_place, fastestLap, fastestLapSpeed, driverId, rank, grid, age\_as\_of\_race, statusId, laps.
* Split dataset into train (1950-2010) and test (2011- ) based on race\_year
* Fit random forest model
* Find model stats **(mse)**, select the best model,  and most important feature **(“points”)**



The model has a mse of 0.0014 and R^2 of 0.984, which means the random forest model is relatively good at predicting the result using the above features.



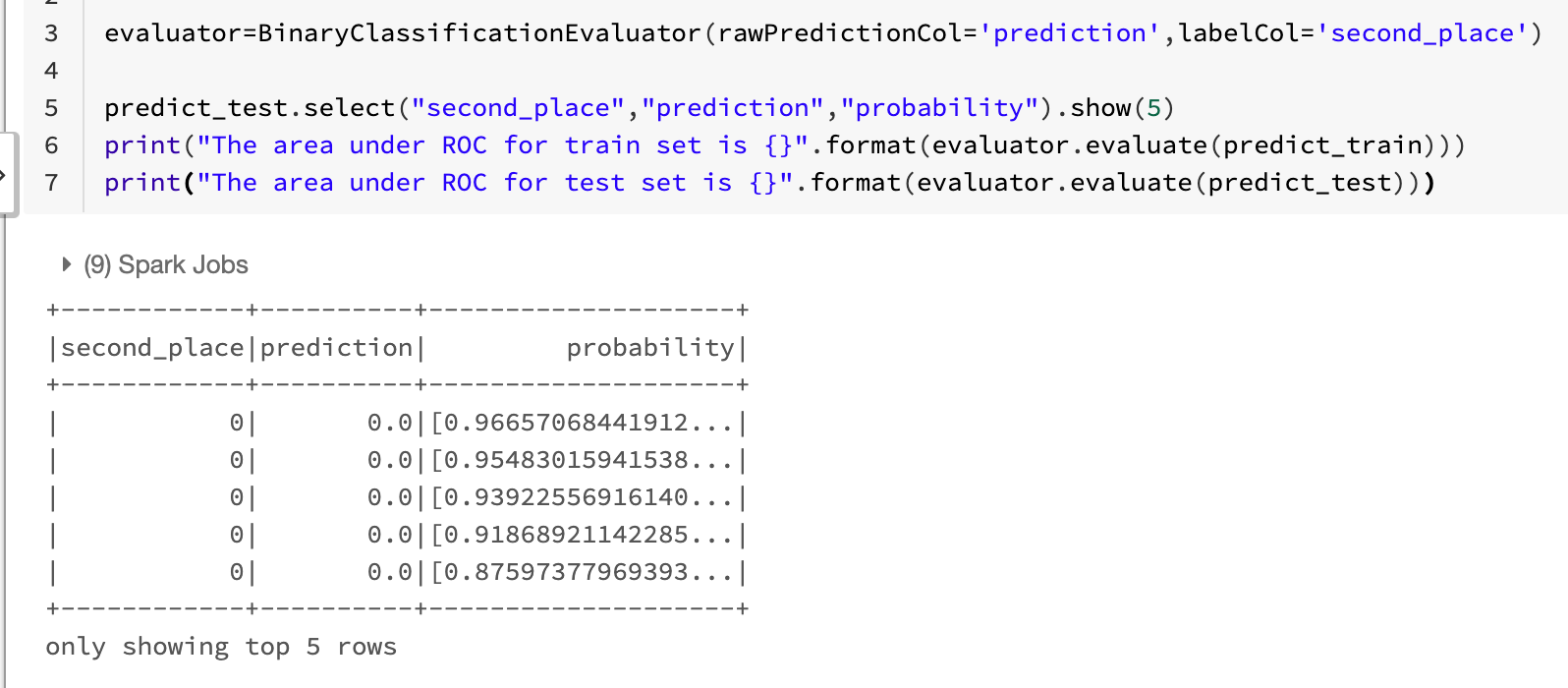


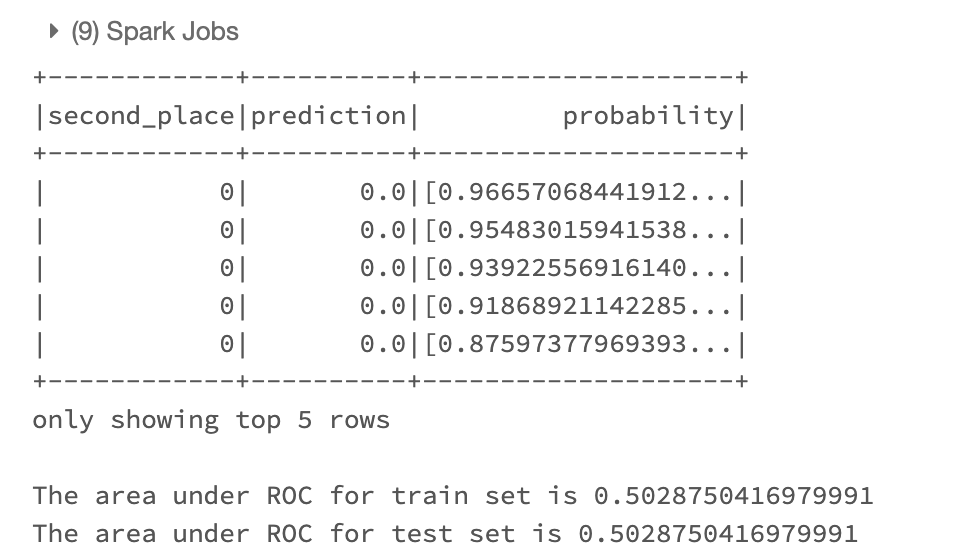
These above are the screenshots from all models and the best model runs. I ran 5 models and selected the best model based on its MSE and R^2.



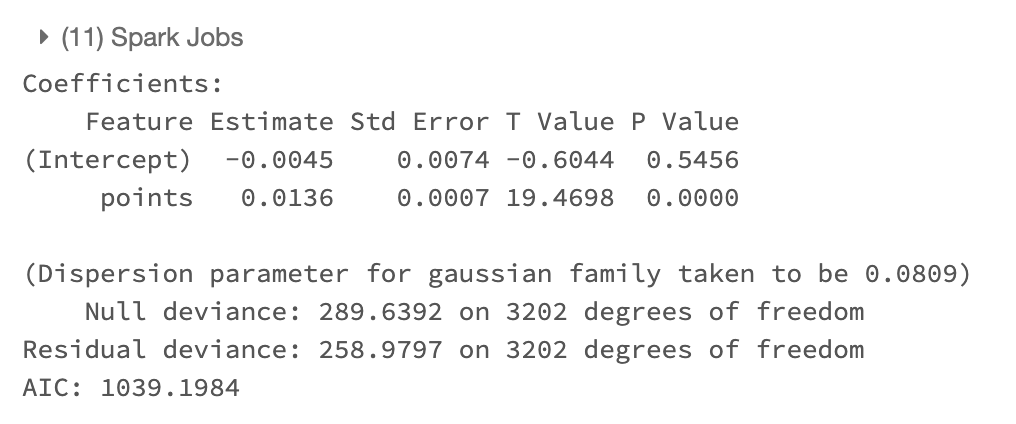
Feature Validation:

* Run logistic regression and generalized logistic regression
* Assess metrics e.g. coefficient, p-value, t-value, and R^2



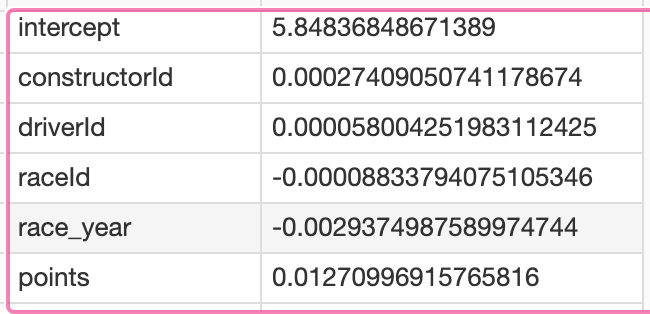


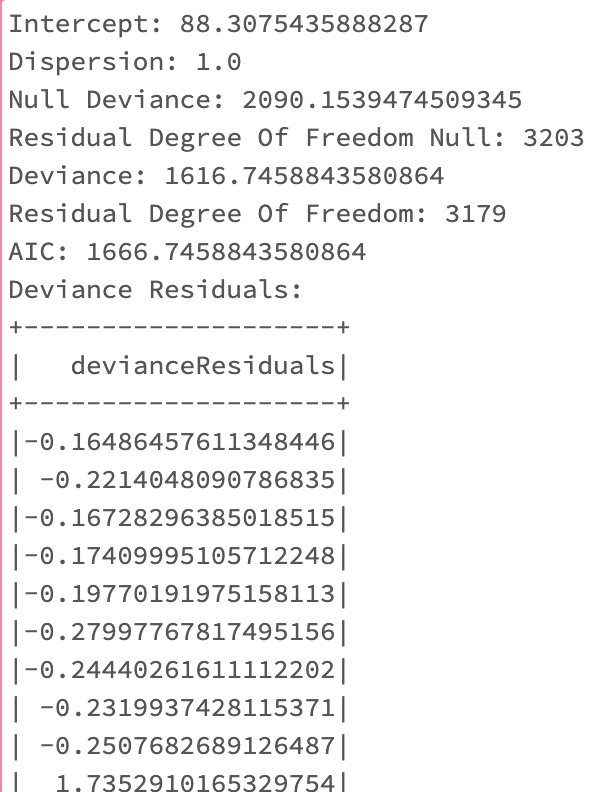
(prediction above is for Points)



(prediction above is for Points)

From the output of “points”, we can see that the variable has a predictive power (0.0136 as coefficient and p-value =0, and ROC is 45%, which has discrimination but in general still a reasonable model.





From the output of all variables, we see “points’ has a coefficient of 0.0127, and deviance residuals has decreased by added variables (e.g. the null deviance decreased from 2090 to 1616). This means that the added variable does have a predictive effect and is a good predictor.

From the above analysis, I found the result is slightly different from Q1 inference model predicts, which derives that “grid\_3” is the most important feature, with coefficient of 3.63 and is statistically significant from p value of close to 0. I think the difference comes from: We used different model for prediction. While the inference used logistic regression for feature validation, the prediction uses random forest model to avoid overfitting. In addition, in terms of interpretation of the result – the inference model makes sense in terms of predicting that gird\_3 yield the best 2nd place prediction. It might be that grid 3 is closer to the inner circle, and might have a slight advantage in the first few runs, and therefore can better give driver confidence and achieve good ranking in the end. For prediction model, it found out that in Q2, points is the best predictor for 2nd place, with p value close to 0, and each additional points contributes to e^0.013 in terms of predictive power. It is revealing of the fact that past performance might be a good indicator of future performance. For example, if someone has a good track record and has a good past performance, he/she might be more likely to have a 2nd place ranking in the upcoming race, which is very promising outcome.

From the overall inference and prediction, it is recommended that driver can improve their overall performance so that they are likely to achieve a higher ranking (2nd place) in the next race. It will also help if they start from grid\_3 rather than from other grids, since they are more likely to have a higher ranking (2nd place) if they start from there.