1. Explanatory Modeling

- describe your model, and explain why each feature was selected

After simple exploratory data analysis, I choose several features that have obvious differences in second place and other places.

1. points: points of first place is the highest, and the points of second place is much higher than drivers at other positions
2. milliseconds: This feature describes the time driver use to finish the race. The less the milliseconds, the higher chance to arrive at first or second.
3. rank: This variable is a rank of time to finish the fastest lap.
4. constructorId: The type of constructor installed on the race car might be a factor affected the race result.
5. fastestLapSpeed: The fastestLapSpeed is an indicator that drivers' ability to increase speed.
6. statusId: This variable represents the status of driver in the race, like if finish the race or if there are some accidents.
7. constructor wins: The number of wins of each kind of constructor used in the race car.
8. age: I suppose the result is related to age. But I am not sure if younger drivers are braver to win the race or older drivers have more experience to win.

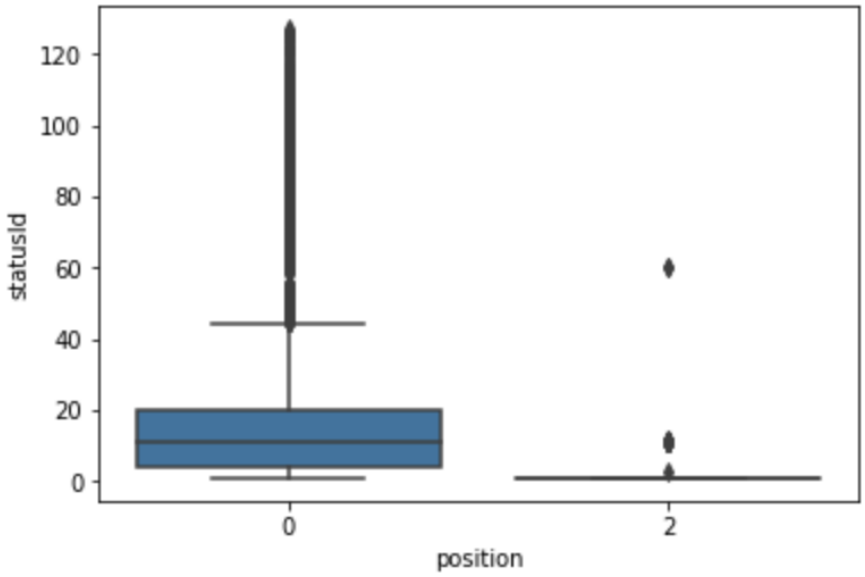
- provide statistics that show how well the model fits the data

```
                           Logit Regression Results
==============================================================================
Dep. Variable:                position   No. Observations:                20773
Model:                           Logit   Df Residuals:                    20765
Method:                            MLE   Df Model:                            7
Date:                 Fri, 08 May 2020   Pseudo R-squ.:                  0.3518
Time:                         19:35:12   Log-Likelihood:                -2296.5
converged:                        True   LL-Null:                       -3542.8
                                         LLR p-value:                     0.000
==============================================================================
                     coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
points               0.2243      0.013     17.327      0.000       0.199       0.250
milliseconds      8.553e-08   2.36e-08      3.618      0.000    3.92e-08    1.32e-07
rank                -0.0195      0.011     -1.756      0.079      -0.041       0.002
constructorId       -0.0028      0.001     -2.919      0.004      -0.005      -0.001
fastestLapSpeed     -0.0155      0.001    -13.147      0.000      -0.018      -0.013
statusId            -0.3473      0.026    -13.308      0.000      -0.398      -0.296
constructor_wins     0.0216      0.017      1.280      0.201      -0.011       0.055
age                  0.0106      0.008      1.354      0.176      -0.005       0.026
==============================================================================

Possibly complete quasi-separation: A fraction 0.26 of observations can be
perfectly predicted. This might indicate that there is complete
quasi-separation. In this case some parameters will not be identified.

Command took 0.28 seconds -- by xh2434@columbia.edu at 5/8/2020, 3:35:12 PM on My Cluster
```

With Logit method in statsmodels.api, I create a logit model to explore the interactions between arriving at the second place and the factors I choose. As you can see in the table, Pseudo R squared is 35.18%.

- what is the most important variable in your model? How did you determine that?

According to the summary, the most important variable is statusId. The estimated coefficient is bigger than that of other factors, which is -0.3473. This means the smaller the statusId is, the more possible the driver arrives at second place. I check the status.csv, when statusId is 1, the status is finished. As the number of statusId increase, the status change to various accidents like collision and disqualified.



- provide some marginal effects for the variable that you identified as the most important in the model, and interpret it in the context of F1 races: in other words, give us the story that the data is providing you about drivers that come in second place

```
        Logit Marginal Effects
=====================================
Dep. Variable:                position
Method:                           dydx
At:                           overall
================================================================================
                     dy/dx    std err          z      P>|z|      [0.025      0.975]
--------------------------------------------------------------------------------
points               0.0070      0.000     17.954      0.000       0.006       0.008
milliseconds      2.681e-09   7.41e-10      3.617      0.000    1.23e-09    4.13e-09
rank                -0.0006      0.000     -1.757      0.079      -0.001    7.07e-05
constructorId    -8.893e-05   3.04e-05     -2.921      0.003      -0.000   -2.93e-05
fastestLapSpeed     -0.0005   3.59e-05    -13.569      0.000      -0.001      -0.000
statusId            -0.0109      0.001    -12.585      0.000      -0.013      -0.009
constructor_wins     0.0007      0.001      1.280      0.201      -0.000       0.002
age                  0.0003      0.000      1.354      0.176      -0.000       0.001
================================================================================
```

Command took 0.10 seconds -- by xh2434@columbia.edu at 5/8/2020, 3:38:40 PM on My Cluster

- does it make sense to think of it as an "explanation" for drivers arriving in second place? or is it simply an association we observe in the data?

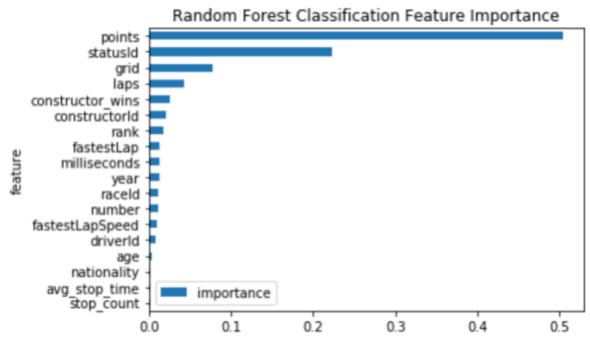
I think it can be an explanation for drivers arriving in second place. Because there is high possibility that various accidents happen and make the drivers fail in the race. Only a small part of drivers can finish the race safely and successfully, and it is a key factor to compete the first or second place.

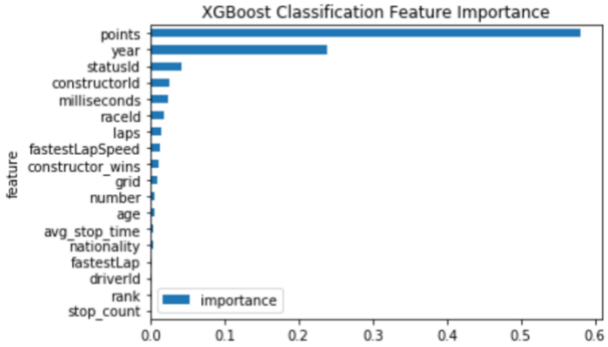
2. Predictive Modeling

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

I select features by algorithms like Random Forest and XGBoost to get the importance of features and pick those with high importance. In Random Forest Model, the most importance features include points, statusId, grid, laps, constructor_wins, constructorId.
In XGBoost Model, the most important features include points, year, statusId, constructorId, milliseconds, raceId.

In conclusion, **points, statusId, constructorId** are the overlap important features under two algorithms.





- 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

Model 1: Random Forest Classification. The accuracy score of this model to predict data between 2011 and 2017 is 75.1%. From the classification report, we can find that the precision of predicting second place is 0.99, but the recall ratio is only 0.51. This means the model exist too many False Negatives, and the model predict second places as other places.

```
Accuracy Score of Random Forest Classification: 0.7510474860335196
ROC AUC Score of Random Forest Classification: 0.7510474860335195
              precision    recall  f1-score   support

           0       0.67      0.99      0.80      2864
           2       0.99      0.51      0.67      2864

    accuracy                           0.75      5728
   macro avg       0.83      0.75      0.74      5728
weighted avg       0.83      0.75      0.74      5728
```

Model 2: The accuracy of XGBoost Model is 99.93%. Combined with the metrics like ROC-AUC Score and classification report, I think this model can predict the second place accurately.

```
Accuracy Score of XGBoost Classification: 0.9993335554815062
ROC AUC Score of XGBoost Classification: 0.9961757839579172
              precision    recall  f1-score   support

           0       1.00      1.00      1.00      2864
           2       0.99      0.99      0.99       137

    accuracy                           1.00      3001
   macro avg       1.00      1.00      1.00      3001
weighted avg       1.00      1.00      1.00      3001
```

Model 3: The accuracy score of KNN Model is 66.2%. From the classification report, we can say it has the same problem with random forest and the problem is worse in KNN model.

Although the model will not classify other places as second places, it cannot distinguish all second places from the individuals.

```
Accuracy Score of KNN Classification: 0.6620111731843575
ROC AUC Score of KNN Classification: 0.6620111731843574
              precision    recall  f1-score   support

           0       0.60      0.98      0.74      2864
           2       0.95      0.34      0.50      2864

    accuracy                           0.66      5728
   macro avg       0.78      0.66      0.62      5728
weighted avg       0.78      0.66      0.62      5728
```

- 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?

I provide the feature importance from Random Forest model and XGBoost Model. In Random Forest Model, the most important feature is points and the second important feature is statusId. In XGBoost Model, the most important feature is points and statusId is the third important feature. However, XGBoost has a better predicting performance although it doesn't put the most important explanatory variable "statusId" in (1) as the most important prediction factor. Prediction model is different from explanatory model. And there are many factors that are associated with the outcome, but they don't have causal relations.

feature	importance
year	0.012370
raceId	0.012010
driverId	0.008119
constructorId	0.020416
number	0.011812
grid	0.077829
points	0.503553
laps	0.043453
milliseconds	0.013085
fastestLap	0.013252
rank	0.017825
fastestLapSpeed	0.009241
statusId	0.222736
stop_count	0.001248
avg_stop_time	0.001766
constructor_wins	0.025263
nationality	0.002636
age	0.003387

Random Forest Model

feature	importance
year	0.238525
raceId	0.018425
driverId	0.002426
constructorId	0.025573
number	0.005153
grid	0.009455
points	0.579822
laps	0.014715
milliseconds	0.023987
fastestLap	0.002623
rank	0.002411
fastestLapSpeed	0.012580
statusId	0.042117
stop_count	0.000000
avg_stop_time	0.003386
constructor_wins	0.010546
nationality	0.003354
age	0.004902

XGBoost Model