(3) [25pts] This task is inferential. You are going to try to explain why a constructor wins a season between 1950 and 2010. Fit a model using features that make theoretical sense to describe F1 racing between 1950 and 2010. Clean the data, and transform it as necessary, including dealing with missing data. [Remember, this will almost necessarily be an overfit model where variables are selected because they make sense to explain F1 races between 1950 and 2010, and not based on algorithmic feature selection]

From your fitted model:

- describe your model, and explain why each feature was selected
- provide statistics that show how well the model fits the data
- what is the most important variable in your model? How did you determine that?
- 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 constructors that win seasons
- does it make sense to think of it as an "explanation" for why a constructor wins a season? or is it simply an association we observe in the data?

Data Preprocessing:

1. Merge preprocessed data that describe features of a race with the status of each race. The status is a binary variable, where if wins equals to 1, the driver wins the race.

2. Calculated the average age or the driver for each constructor in each season.

3. Created a season variable to account for the role of seasonality in F1 race by extracting and grouping date of the race. One-hot-encode the four seasons to fit the model.

4. Created a continent variable to account for constructor's nationality. This is a continuous variable where 1,2,3, and 4 each means if the constructor is from Europe, North and South America, Asia, and Africa.

5. Grouped by year and constructor ID and Calculated the average of each explanatory variable for every constructor in every year between 1950 and 2010.

Model Building:

1. Run a stats model Linear Regression model to explain why a constructor wins a season by looking at the variable's coefficients and statistical significance.
2. Build a machine learning linear regression model to look at model evaluation metrics and determine if our model is overfitting
3. Feature selection: race_year, points, constructorId, grid, age_as_of_race, positionOrder, points_x, s_autumn, s_spring, s_summer, s_winter.
4. Split dataset into training and test set
5. Fit linear regression models
6. Find model stats **(rmse)**, select the best model, and most important feature **("points")**

Model Result & Interpretation:

-- Stats model Result Table --

OLS Regression Results

Dep. Variable:	wins	R-squared:	0.609
Model:	OLS	Adj. R-squared:	0.605
Method:	Least Squares	F-statistic:	156.3
Date:	Sat, 10 Apr 2021	Prob (F-statistic):	2.60e-177
Time:	00:12:59	Log-Likelihood:	-355.12
No. Observations:	912	AIC:	730.2
Df Residuals:	902	BIC:	778.4
Df Model:	9		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	-0.0357	0.085	-0.419	0.676	-0.203	0.132
grid	-0.0063	0.003	-2.358	0.019	-0.011	-0.001
positionOrder	0.0084	0.004	2.049	0.041	0.000	0.016
points_x	0.3000	0.012	25.998	0.000	0.277	0.323
laps	-0.0006	0.001	-1.158	0.247	-0.002	0.000
age_as_of_race	0.0025	0.003	0.867	0.386	-0.003	0.008
s_autumn	0.0517	0.066	0.787	0.431	-0.077	0.181
s_spring	-0.0363	0.060	-0.606	0.545	-0.154	0.081
s_summer	-0.0092	0.048	-0.193	0.847	-0.103	0.084
s_winter	-0.0419	0.064	-0.653	0.514	-0.168	0.084
continent	-0.0577	0.034	-1.722	0.085	-0.124	0.008

Omnibus:	269.754	Durbin-Watson:	1.976
Prob(Omnibus):	0.000	Jarque-Bera (JB):	4366.285
Skew:	0.897	Prob(JB):	0.00
Kurtosis:	13.568	Cond. No.	1.86e+17

Warnings:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
[2] The smallest eigenvalue is 1.38e-28. This might indicate that there are
strong multicollinearity problems or that the design matrix is singular.

- **R-squared** is 0.609, indicating the logistic regression accounts for **61%** of the variance. The model fits the data somewhat well.

- **Grid, PositionOrder**, and **points** are all **statistically significant** features. Points have the highest coefficient, which makes sense given it is used to determine the outcome of both

the Drivers' and Constructor's World Championships. Each season, the Championship is awarded to the driver and the team that has scored the highest number of points.

- **Interpretation:** Holding all other variables at fixed values, we will see a **30%** increase in the **odds** of the constructor winning the season for every extra **point** the constructor earns. Points are awarded based on the position the driver finishes at.

-- Machine Learning model Result Table --

```
rmse: 0.545154665782234
rmse_train: 0.5716539710859395
mae: 0.338627292372811
R2: -4.172972188154489e-05
```

- Comparing the RMSE score on the training set and the test set, we can see that the RMSE score fitted on the training set is higher than that fitted on the test set. Therefore, our model is overfitting as expected.

Date: 2021-04-17 22:48:31 Source: Q3_Inference_Model

Duration: 4.3s Status: FINISHED

▾ Notes

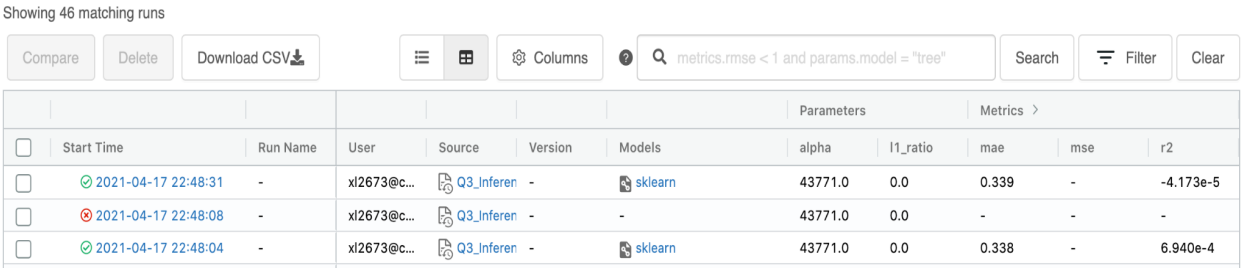
None

▾ Parameters

Name	Value
alpha	43771.0
l1_ratio	0.0

▾ Metrics

Name	Value
mae	0.339
r2	-4.173e-5
rmse	0.545



Overall, by looking at the inference model, it's clear that only points, grid, and PositionOrder are statistically significant variables at explaining why a constructor might win a season. And by looking at the coefficient, we can find the total points earned by the constructor in each season has the highest explanatory power on the race result.

Furthermore, by looking at the linear regression machine learning model, we can clearly detect the overfitting issue in our model, because the RMSE score fitted on our training set is slightly higher than that fitted on the test set. This is expected because we want to include all variables that help to explain why a constructor wins a season, which means the model tries to explain more idiosyncrasies in the F1 dataset.