F1 Capstone Presentation Script

Project Objective

Background

- Formula 1 is the highest class of open wheel racing in the world
- The World Driver's Championship has been contested since 1950 and the sport has evolved into a commercial juggernaut, with billions exchanged for TV rights and sponsorship for teams, known as constructors
- The 2021 title fight between Lewis Hamilton and Max Verstappen means viewers are at their highest ever levels globally, with Sky attracting 2.3 million viewers in the UK for races, the highest ever audiences for pay per view broadcasters

Driver prediction importance

- Predicting the result of a race consistently is almost impossible thanks to multiple external factors
- However, there is also importance in predictions for if a driver moves
- Making them could benefit teams to see if results are mirroring an underperforming driver who has left in the past, and whether they should be making a similar decision as well and there are also betting markets for driver transfers
- It is worth considering that drivers are kept for more than just driver merit, such as pay drivers whose families bring in development money to smaller budget teams
- However, this level of technical and political data is not readily available and so this will not be studied

Data

Data Target

- From original csv files, I wanted driver data compiled for a season of racing, with attributes such as total points, final position, average position and total races through the year
- Year by year data with the drivers that raced were concatenated, with Year and DriverID columns effectively forming a composite key
- Finally, there would be the driver_move column, the classifier and binary target to predict which at the moment is unknown and I will discuss on slide 8

Data observations

- F1 in the 20th century had less uniform entry requirements, often times teams showing up for one or two races a season, leading to 108 racers in one season
- In 2021, only 21 drivers competed across the 22 races
- As a result of this I took data for only the last 20 years from 2002

Driver_move

- No given value actually existed for driver_move which is equivalent to the driver leaving a team or a churn value
- For each season I took the modal constructor that the driver raced for, occasionally drivers switch seats mid-season due to performance

- The value of 1 or 0 was based on finding duplicate driver and constructor ID pairs in consecutive years and returning that result, and I then inputted 2021 data manually as this is known but there are no races in 2022. There are errors of this which I will talk about in limitations on slide 19
- This classifies our 481 data entries into the following categories. As there is roughly a 50:50 split, I will treat accuracy as the key metric to measure machine learning

Model generation and tuning

- Used two main focuses classifier models for this project

Logistic regression

- Logistic regression is used when looking for a discrete binary target for an input to predict based on prior data observations
- It is used to estimate the probabilities of events and generate a boundary of probabilities based on data features on which a decision can be made
- A cut off value is then used to generate a statistical classification
- Data manipulation is needed if continuous data not linear, I will clarify this on slide 13

Decision tree and random forest

- A decision tree represents a series of sequential options taken to reach a result, each decision node branching the data, until end nodes or leaves are reached representing a final decision for that group of data points
- A random forest is similar but has multiple tree based-decision outcomes, combines them and generates a final output
- Both of these methods have a negative of overfitting the data if too many levels are introduced
- Next is about deciding which attribute columns are to be used

Correlation Heatmap

- The following heatmap shows where generated rows have correlation with one another,
- We generally expect to see this when categories are similar, however this it's important to select no correlating columns when building a model
    - (Each variable should be independent in its own right
    - Coefficient estimates can swing wildly based on small changes
    - Reduced precision of estimate)

Discrete binary variables

- These graphs show the count of two binary attributes by driver remain and driver leave
- The left shows an example of a used attribute showing large disparity, the right an unsued one with low disparity

Continuous variables

- This histogram shows the driver_move value as blue and orange colours and the count of each for the drivers median finishing position in that season
- It is clear that position number less than 6 and higher than 14 show large disparity, but it is less clear for middle positions. This is where decision trees may outperform logistic regression without intense feature engineering

Model outcomes

Model scores

- Here is a list of scores from the models run and the improvement from the baseline starting position. As data is almost evenly split, 50% accuracy represents complete guessing
- The final accuracy of the best random forest model was 78.3% on training, 72.7% on test showing a little overfitting but not to a detrimental effect
- One of the models also showed higher test accuracy than training, however this most likely due to the small data set
- The final model also showed both strong precision and recall

Matrix and final feature importance

- This is the decision matrix on the test set of that model
- There is also the graph on the right that shows the feature importance of specific attributes used in the final model, showing median position and percentage of constructor points as key classifiers
- Eliminating less used columns can be key to reducing the overfitting problem of decision trees, as well as cross validation that was used with all models

Conclusions

Successes

- Me personally achieving a final 72.7% accuracy I was very happy with after aiming for 75%
- I had no idea if this was something that could be predicting base on the limited data I had and also solving an unknown problem that doesn't appear to have any small scale attempts done by individuals

Limitations

- However, I do accept that this project has limitations
- We are making predictions based on a whole season of results, sometimes drivers are confirmed at this point, but generally this can happen anywhere from 60 to 90% of season completion
- There is a limited source of data to model on, even if every season/driver pair was included, this would only be a few thousand results
- Point system alterations can affect driver metric and final championship position and now with sprint qualifying awarding points
- Leave value does not account for a new team name or branding in the data, so the driver may have been 'retained' but the driver_move column can not account for this

Improvements

- Identify team name changes
- Include further metrics on which to measure drivers success
- Also having a categorical output, for driver promotion, retirement or relegation on which to potentially perform a clustering project
- Finally I would just like to thank Lisa and Blair for all of their help and assistance during the training for and completion of my project
- Thank you very much and very quickly are there any questions