CKME 136 – Final Report

## A Study on Laptimes at the Nürburgring

**Supervisor: Professor Tamer Abdou**

**Samuel Tang, 501 012 417**

A Study on Laptimes at the Nürburgring

# Introduction

As a car enthusiast, I am often looking for the latest trends in motorsport engineering. I have always wanted to combine my passion for data with motorsport. That is why I have chosen this “Top 100 Nürburgring” dataset. The dataset and problem that I am targeting to solve is a model to predict lap times at the famous Nordschleife – Nürburgring. This model will be built using Chris Scott’s Kaggle dataset, which features the top 100 laptimes for cars up to the model year 2017. Some of the determinants recorded are Model Year, Car Make, Car Model, and Lap Time. Other features such as Weight, Engine Placement, Horsepower, Transmission, Temperature, etc. will be considered and added.

By creating this model, I hope to discern what are the greatest influencers to laptime and how car manufacturers can better improve their cars. Another hope is to create an accurate lap time predictor when given as much information as possible with these indicators.

Main Research Question/Problem: What variables are the greatest influencers upon laptime?

A link to the GitHub Repository: https://github.com/samuel-adrian-tang/CKME-136

# Literature Review

Several searches and queries were made to try and find relevant research concerning automotive performance; however, ALL searches have yielded zero results. The searches were done through JSTOR, Google Scholar, and through other search engines. The only sources of information appear to be informal and not a result of academia.

The sources that I have consulted are automotive journaling magazines as well as interviews from various racing drivers. While no formal scientific research (concerning this topic) has been published or made publicly available, there is an overwhelming consensus from these field “experts”.

There are five major external influencers to lap time:

1. Tyre – How much contact surface area is available, and how does it perform when hot? How much lateral force can the tyre sustain?
2. Engine – Horsepower, torque, and other engine characteristics.
3. Aerodynamics – How much downforce is the car able to generate at \_\_\_ speed?
4. Weight – How much does the automobile weigh?
5. Brakes – How much stopping force is available? How does performance suffer when hot?

All five of these metrics are scientifically measurable. However, the data is simply not available for brake performance, aerodynamic pressure, or tyre performance. It cannot be understated that racing is a form of sport; any scientific measurement applied to sport must be taken with a grain of salt.

There are internal influencers as well:

1. Driver confidence – How familiar is the driver with the car/instrument?
2. Driver skill – The ability to place the car at optimal entry and exit points (of a turn).
3. Track familiarity – How familiar is the driver with the track?

Regarding these three factors, we can disregard the third statement as all laps were done at a single track. Driver confidence and driver skill are also negated in this instance as manufacturers would have their own test drivers (peak confidence, skill, and familiarity), race the cars.

# Dataset

The original dataset is from Kaggle (<https://www.kaggle.com/scottdchris/nurburgring100/kernels>), and featured five variables (Position, Time, Year, Make, Model). The data featured 100 entries and was likely scraped either from a manufacturer website, or a third party site such as <https://fastestlaps.com/tracks/nordschleife>.

I have removed several entries from the dataset; they will not be considered for these reasons:

* Small manufacturers
* Specialty manufacturers
* Not road/street legal
* Multiples (unless substantial differences)
* Modified Vehicles

I have also added 13 measurable and accurate variables that will hopefully allow us to investigate this problem even further. In total, there are now 81 entries and 18 variables, and zero missing data points.

In the future, I may add other features to this dataset such as brake material, tyre, weather, and maximum lateral force. Mins2Secs was created to convert the recorded laptimes into seconds, and KG/HP is simply a function of Weight divided by Horsepower.

There have been some preliminary observations of the data set:

* Most cars are of the Coupe body type, the others being Sedan and Hatchback
* Lighter cars appear to have greater torque, manual transmissions and better laptimes
* AWD systems appear to perform quicker than their RWD and FWD counterparts
* Hybrids are featured as both the quickest and slowest but also heaviest cars
* As expected, Horsepower and Torque are heavily correlated, they also correlate with superior laptimes

# Approach

## Step 1: Find Dataset

A search was done on Google and Kaggle to find the relevant dataset as well as any other assisting data.

## Step 2: Edit Dataset

The dataset was then edited to find supplemental information as well as remove unnecessary information. There were no missing values or incorrect values to report. Categorical variables will be converted to binary variables for analysis.

## Step 3: Preliminary/Exploratory Analysis

A brief look at the complete dataset will be undertaken. Any explicit and noticeable trends will be noted as well as a basic analysis of correlation, outliers, and basic summary statistics.

## Step 4: Building the Model

Multiple models and methods will be done at this stage.

1. Feature selection – dimensionality reduction via correlation analysis, variance analysis, stepwise regression, and etc.
2. Final variables will be regressed upon Mins2Secs
3. Classification/clustering, the cars will be categorized into different clusters, which will assist in the regression accuracy

## Step 5: Testing and Tuning the Model

These models would then be evaluated on a variety of different metrics (such as R2, RMSE, MAE, AIC and BIC). Also, I have scaled certain variables to allow for better modeling.

## Step 6: Report and Visualization

At this stage, we will beautifully visualize our findings using Tableau. The report will be a concise business style summary that highlights all the key findings and numbers.

# Results

After analysing the data, we have determined that seven of our variables influence a car’s laptime. The seven variables are:

1. Year
2. Weight
3. Horsepower
4. Torque
5. Cylinders
6. AWD
7. Mid-Engine

Scaling the variables show that increasing horsepower is the most effective thing to focus on to decrease laptime. One of the most surprising results is that Year (in which the car is manufactured) is also a major factor in decreasing laptime. In this case, Year would capture many technological advances that we are currently unable to represent in an individual variable. Newer cars tend to have stronger, lighter, and more efficient materials, it is just simply too difficult to account for those factors apart from the “Year” variable.

One of the best features that we added to the model was the addition of clusters. These clusters placed the cars within four different groups that were primarily determined by laptime. Now, this may seem as too strong a handicap, especially since the model is supposed to predict laptime. However, it makes intuitive sense that the cars would be separated that way as manufacturers would benchmark their cars against others using laptime as a main benchmarking metric!

The result of adding these clusters increased the accuracy of our model by a significant margin! On all measurable metrics, there were markable improvements.

|  |  |  |
| --- | --- | --- |
| Metric | Without Clusters | With Clusters |
| RMSE | 8.3615 | 3.8043 |
| R-Squared / Adjusted R-Squared | .6398  .6053 | .9264  .9159 |
| Mean Absolute Error | 14.96942 | 7.20856 |
| Akaike Information Criterion | 414.6986 | 327.594 |
| Bayesian Information Criterion | 432.9268 | 351.8983 |

# Conclusions

To conclude, decreasing weight, increasing horsepower, and adding an All-wheel-drive system is the best way to better your car’s laptime. As evidenced by many of the examples, it is easy to compensate a shortcoming with a strength in another variable.

In researching this study, I can discern many shortcomings to my analysis.

* Racetracks play to different strengths, findings from this study may not be applicable to laptimes in general or at different circuits
* Many important factors were forced to be captured by “Year” as we were unable to isolate different variables such as aerodynamics, differential systems, tyres, or brakes.
* I have tried to keep measurements standard across all cars, however manufacturers and governments have different reporting standards on variables such as weight, torque, and horsepower