D22_report

KAGGLE---FORMULA-1

Team members: Karl Gustav Loog, Silver Vaino

Repository link: here

Business understanding

Identifying your business goals

Background:

For our first Data-Mining project we wanted to choose a topic that was relatively logical and close to us. Logical so that it would fundamentally make sense to data-mine in that field. And close to us so that if we would find ourselves in a difficult situation, where there seems to be no happy endings, then we will find the motivation to keep on going. So we chose Formula 1. Formula 1 is the pinnacle of motorsport. It is the sport where innovation, technology and talent collide. So it has been for years.

Business goals:

Formula 1, as a motorsport, has always been about the data. The car that sits under the driver has a lot of components, a lot of functions. Each of these components and functions has to be pushed to the limit in order to get the best possible result. That could be the best top speed, best aerodynamics, best overtaking speed, or overall best lap time. After each race, each session, the team runs analysis and collects data to improve those aspects previously mentioned. Data is information and information is power, in this case speed.
In this project we want to go back in time and see if we can find different patterns. Such as where do drivers (or cars) tend to DNF (did not finish) and what may have caused it (is there a pattern worth looking into?). Maybe the racetrack or the weather in that

particular area? Or are the more experienced constructors (teams, such as ferrari, alfa romeo, mclaren, and so on...) more successful? These are the problems we try to face. Because this is a wide topic and we have data all the way back from 1950 (when the first official grand prix was held), we can't say for sure if we will finish those problems. Maybe we find something else, something more significant, something that hasn't been discovered yet. Only time will tell.

Business success criteria:

As was said in the Business goals point, at the moment we have two goals. To calculate the probability of the driver (or car) getting a DNF and find out if experienced constructors are more successful. From 1950 to 2017. If we manage to successfully complete these tasks, then we would be satisfied, BUT if we manage to find another task in the process, then we, as sport fans, are probably going to be more satisfied. Therefore overcoming our expectations is the main criteria in measuring our success.

Assessing your situation

Inventory of resources:

- **Team members:** Karl Gustav Loog and Silver Vaino
- **Data:** Race results, Championship results, Circuit information, different statuses (1950-2017)
- **Hardware:** Two computers running windows 10
- **Software:** Python

Requirements, assumptions, and constraints:

We must present our results (our project) in the poster session on dec 16. 2 pm, and we have acquired the dataset we need.

Risks and contingencies:

These are some of the risks regarding the work flow:

Problem	**Possible solution**
Power outage	Working from a different office or at school
Problems with the hardware	We can use home computers as a backup
Problems with time management	Switch shifts

Terminology:

- Circuit - racetrack (such as Monaco)
- DNF - did not finish
- Constructor - team (such as Ferrari, Mclaren, Renault, Williams)
- Different statuses that will be specified when come up

Costs and benefits:

We don’t really pay for anything. Everything is to gain here. We found the database on the internet and we didn’t have to pay for it. Also the hardware has been given to us by the University of Tartu. And all of the software we are probably going to use is either free to use or accessible within the student privileges.

Defining your data-mining goals

Data-mining goals:

Construct a model that indicated the probability of a Diver (or car) having a DNF. And one that indicated whether old constructor are still at their game.
Regardless if we reach them or not, we will present the results in a poster session on dec 16. at 2pm.

Data-mining success criteria:

If we manage to construct these models (previously mentioned) by dec 16. then we would be happy.

Data understanding

Gathering data

Outline data requirements:

For the DNF model we're going to use datasets such as seasons.csv, races.csv, results.csv, circuits.csv, and so on. There is a lot to cover from 1950-2017.
For the constructors task we are going to use datasets constructorResults.csv, constructorStandings.csv and constructors.csv (for now).

Verify data availability:

Link to our database: data

Define selection criteria:

We are most likely going to use all of the sets from the previously brought up database. All except drivers.csv, pitStops.csv and qualifying.csv.

Describing data

The database, to our fortune, consists of probably everything we need to know in order to get the promised results. I won't get into details, because there are a few.

Exploring data

Datasets and their structure:

- **circuits.csv** - circuitId,circuitRef,name,location,country,lat,lng,alt,url
- **constructorResults.csv** - constructorResultsId,raceId,constructorId,points,status
- **constructorStandings.csv** - constructorId,constructorRef,name,nationality,url
- **constructors.csv** - constructorStandingsId,raceId,constructorId,points,position,positionText,wins
- **driverStandings.csv** - driverId,driverRef,number,code,forename,surname,dob,nationality,url
- **drivers.csv** - driverStandingsId,raceId,driverId,points,position,positionText,wins
- **lapTimes.csv** - raceId,driverId,lap,position,time,milliseconds
- **pitStops.csv** - raceId,driverId,stop,lap,time,duration,milliseconds
- **qualifying.csv** - qualifyId,raceId,driverId,constructorId,number,position,q1,q2,q3
- **races.csv** - raceId,year,round,circuitId,name,date,time,url
- **results.csv** - resultId,raceId,driverId,constructorId,number,grid,position,positionText,positionOrder,points,laps,time,milliseconds,fastestLap,rank,fastestLapTime,fastestLapSpeed,statusId
- **seasons.csv** - year,url
- **status.csv** - statusId,status

Verifying data quality

Now that I've had some time to analyse the datasets and process my goals, I think that this dataset should fulfil our needs. The only problem with this database is with any unfamiliar database. You really need to understand it in order to work with it sufficiently.

Planning your project

Task	Karl Gustav Loog	Silver Vaino
1. Getting to know the datasets	3 hours	3 hours
2. Discussing and/or Experiment on how to approach the tasks (consulting if necessary)	4 hours	4 hours
3. DNF task	4 - 5 hours	1 hours
4. Experience task	1 hours	4 - 5 hours
5. Final touch and the preparation of the poster	1.5 hours	1.5 hours

Methods and tools:

All the tools that we are going to use are listed in the previous Inventory of resources sub-paragraph. We are most likely going to use something else in addition to these resources. Time will tell which ones.