NCT pass rate analysis and forecast for given car using Data mining methods

Table of Contents

[Introduction 3](#_Toc511081386)

[1. Problem Statement 3](#_Toc511081387)

[2. Data set description 3](#_Toc511081388)

[3. Data preprocessing and formatting 5](#_Toc511081389)

[4. Data mining model construction 7](#_Toc511081390)

[4.1 Classification rules 7](#_Toc511081391)

[5. Models: 8](#_Toc511081392)

[5.1 Naïve Bayes 8](#_Toc511081393)

[Bibliography 12](#_Toc511081394)

# Introduction

Project was made for Data and web mining module. The goal is to predict result of car test – NCT in Ireland.

# Problem Statement

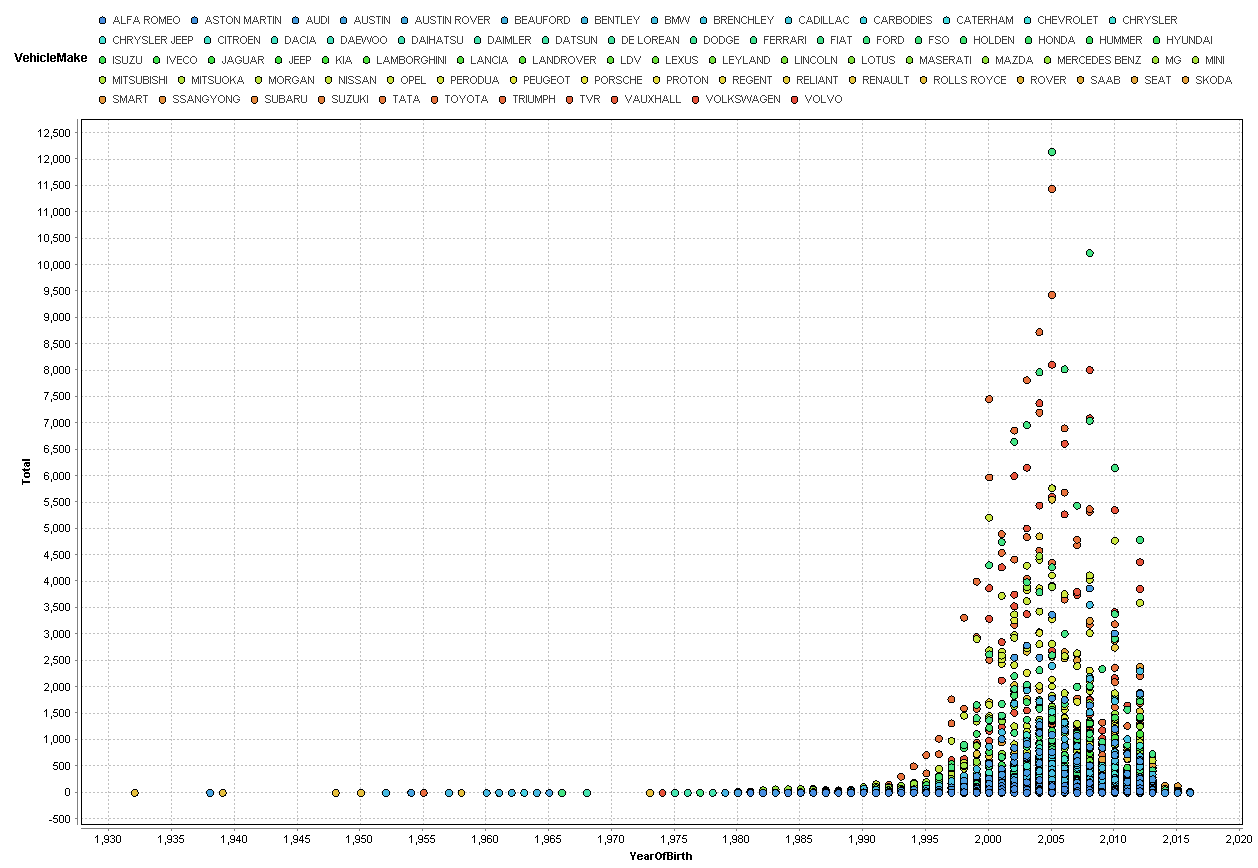
# Data set description

We chose “[2016 Make Model Year Failures at NCT](https://data.gov.ie/dataset/b8e6919f-35ef-4cc8-b694-4cbd5d4e3298/resource/716c8d11-a2b3-4d62-aed3-b12a507c4691)” as a data set for our project. Data set is available to download as a .csv file on data.gov.ie website[[1]](#footnote-1) and statistics come from The Road Safety Authority. It provides anonymized details about all the vehicles that were tested by the National Car Testing Service during 2016.

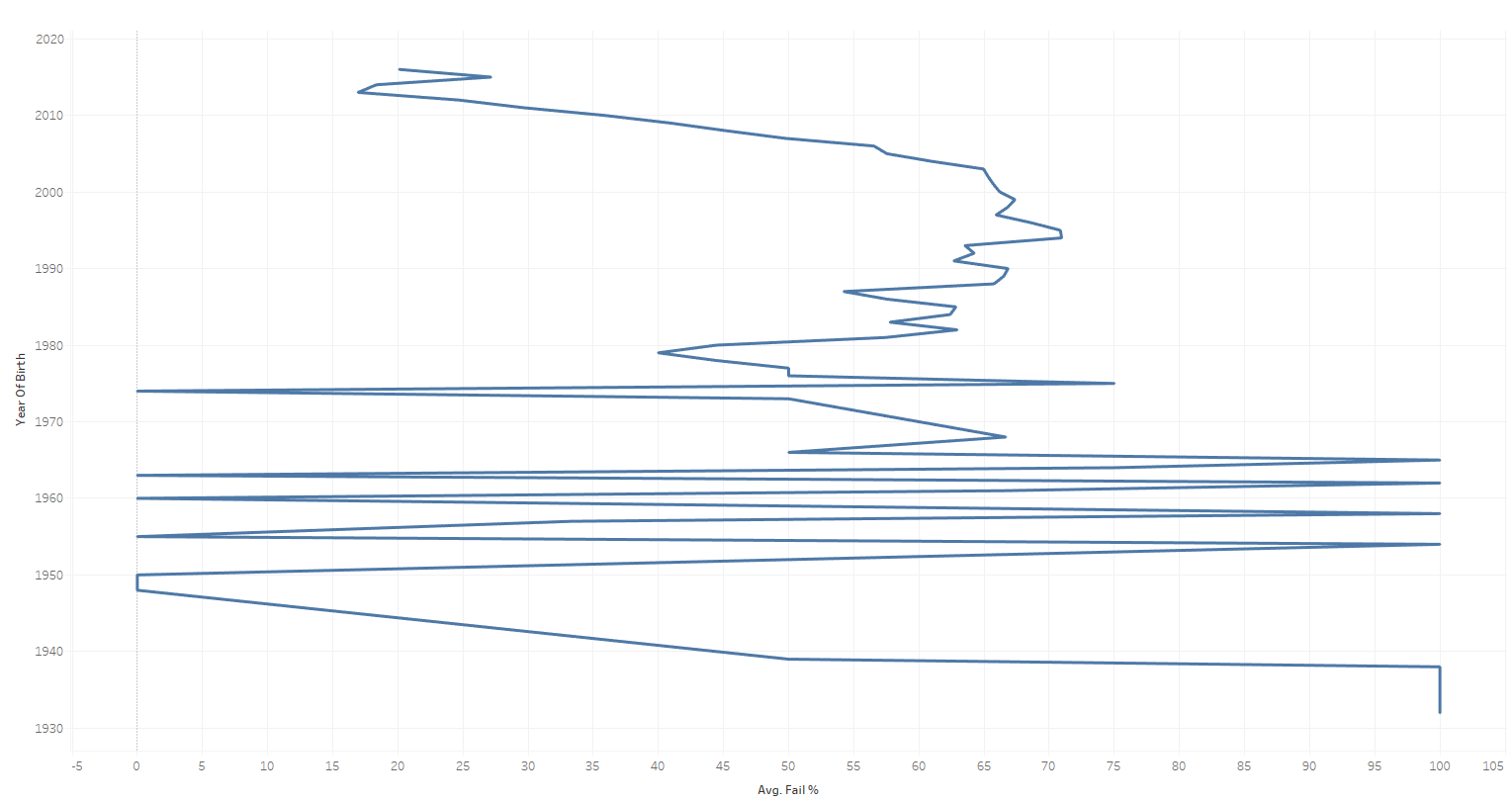
Data set consists of 8089 rows and 37 columns. Data is a mix of a strings and numerical values, columns have the following titles and attributes:

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| VehicleMake | - | string, polynomial | Chassis and Body | - | real |
| VehicleModel | - | string, polynomial | Chassis and Body % | - | ratio, real |
| YearOfBirth | - | date | Side Slip Test | - | real |
| Total | - | range | Side Slip Test % | - | ratio, real |
| PASS | - | real | Suspension Test | - | real |
| PASS % | - | real | Suspension Test % | - | ratio, real |
| FAIL | - | real | Light test | - | real |
| FAIL % | - | real | Light test % | - | ratio, real |
| Vehicle and Safety Equipment | - | real | Brake Test | - | real |
| Vehicle and Safety Equipment % | - | ratio, real | Brake Test % | - | ratio, real |
| Lighting and Electrical | - | real | Emmissions | - | real |
| Lighting and Electrical % | - | ratio, real | Emmissions % | - | ratio, real |
| Steering and Suspension | - | real | OTHER | - | real |
| Steering and Suspension % | - | ratio, real | OTHER % | - | real |
| Braking Equipment | - | real | Incompletable | - | real |
| Braking Equipment % | - | ratio, real | Incompletable % | - | ratio, real |
| Wheels and Tyres | - | real |
| Wheels and Tyres % | - | ratio, real |  |  |  |
| Engine, Noise and Exhaust | - | real |  |  |  |
| Engine, Noise and Exhaust % | - | ratio, real |  |  |  |

We used basic plotting in RapidMiner and Tableau, for initial visualization, to help us understand distribution of data in our data set.



Pic1 Distribution of cars by brand. Created in RapidMiner



Pic2 Distribution of cars by year. Created in Tableau

# Data preprocessing and formatting

We started preprocessing with Python by checking the data for any missing values, duplicates and outliers.

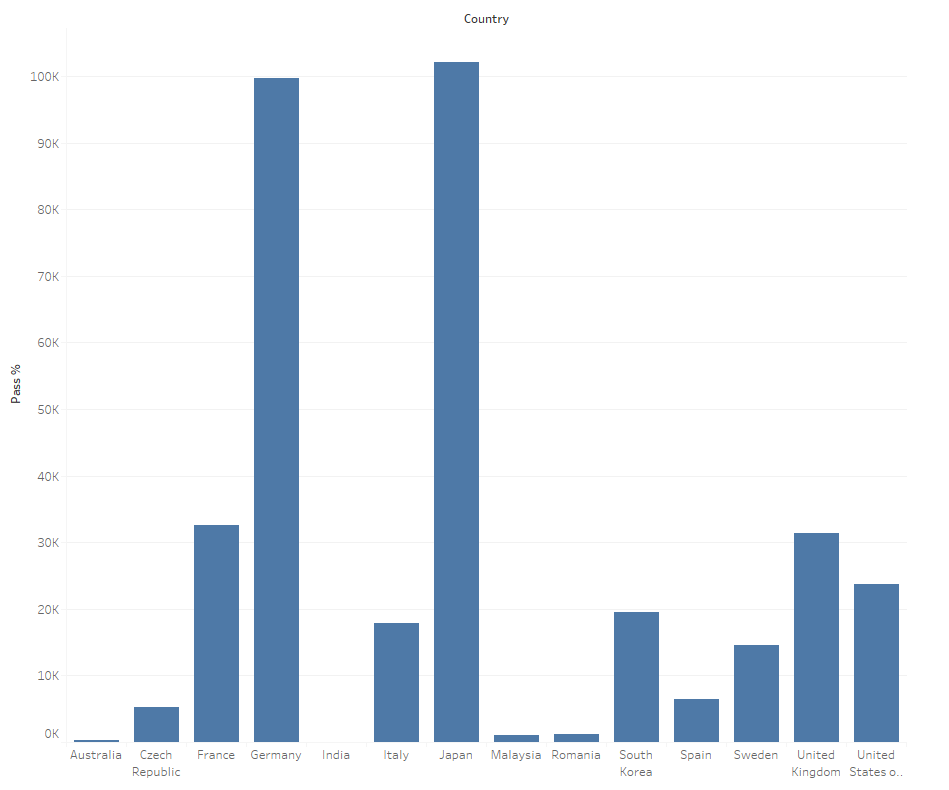
There were no NAN values in our dataset. All duplicates have been carefully checked and any multi entries are the result of normal occurrences of a given value.

Column “ YearofBirth” has been check for consistency of entries. It should contain only year of production and all data seems to be correct. No outliers have been found. For testing purposes we added a column “Country” to our dataset and added the place of origin of tested brands.

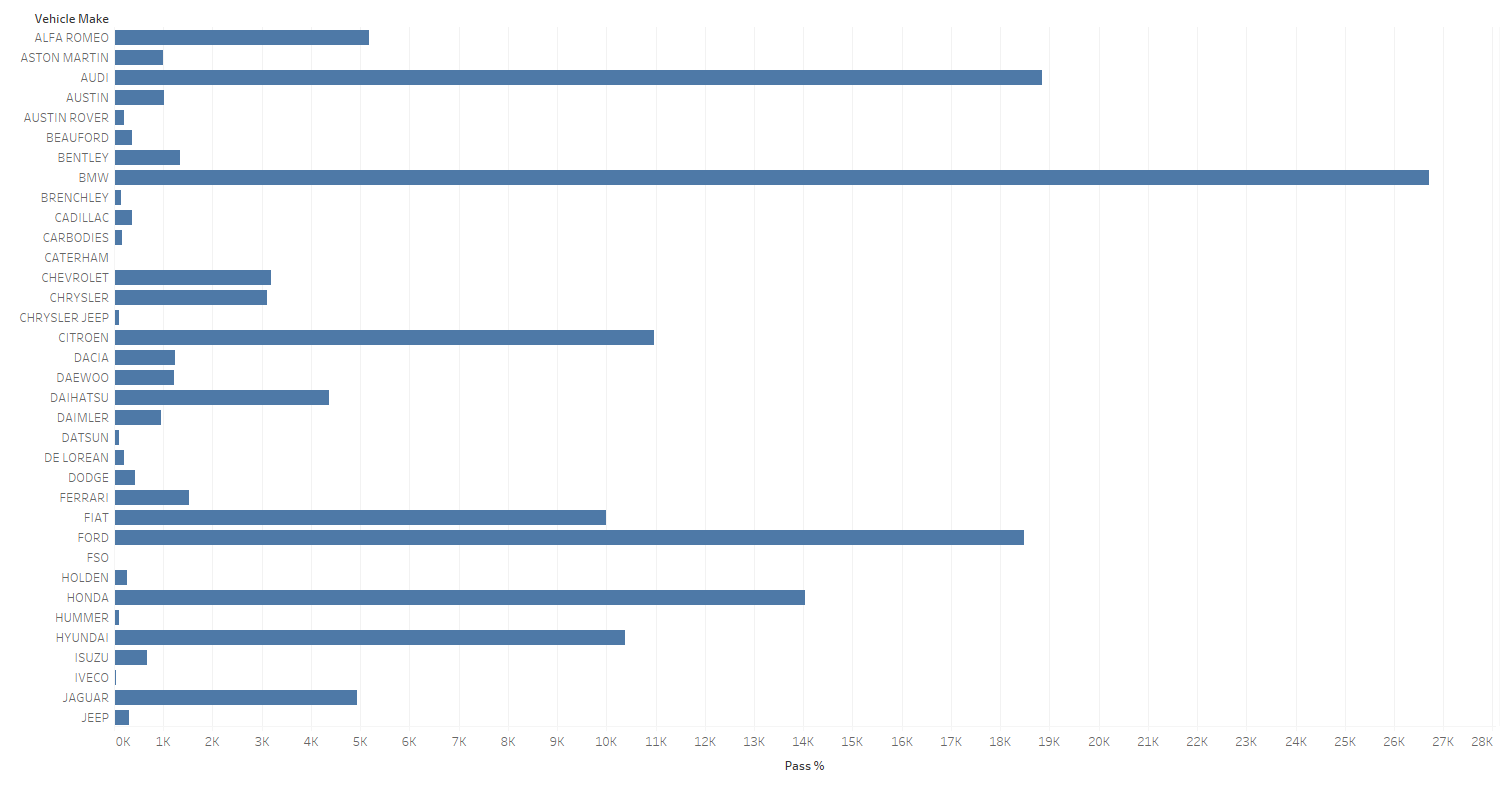
Ratios of failing particular tests are already calculated in our data set, this will be very useful in our model.

As per the Road Safety Authority, all cars manufactured or registered for the first time at least 30 years ago are considered a vintage cars[[2]](#footnote-2) and as such are subject of a different regulations in terms of technical testing. For that reason we decided to limit the data set to the cars registered in or after 1988. They were outliers as proved by visualization on Pic2 on page xx.

We ran some additional visualizations after?? data preprocessing to see more patterns.



Pic3 Distribution of cars by country of origin. Created in Tableau

 Pic4 Distribution of cars by brand. Created in Tableau

# Data mining model construction

## Classification rules

At first we focused on classification rules for our model. The data set contains 19 attributes altogether.

One is a Label – “Pass rate %” - it is our desired outcome. Ideally, we would like to see cars with pass rate above 50%. But this dataset is not typical. 4 attributes are regular polynomial attributes. But another 14 attributes are the potential another outcome, another label. We want to determine the possible reasons of a car failing NCT test.

To apply classification rules we will need all the variables in our model.

NCT test is very specific. It’s different than other tests. If car fail in one of test like for example ‘Wheels and Tires’ test then it fails NCT test.

It is clearly visible that some cars are passing test easier than others. We can see that rules are:

* 2 countries with best pass rate – Japan and Germany
* 6 type of cars passing test easier than the others – BMW, Audi, Ford, Honda, Citroen, Hyundai
* If any of small test is failed all test is failed. So outcome of any small test is our label too.
* ‘Total’ column shows us number of cars tested. We will use that column to set up rule of weighted next column of Pass rate. So for example if only 2 cars have been tested than weighted pass rate is 2. If the 100 cars has been tested than weighted pass rate means is multiple by 100.

Rules has been build based on our work with database, visualization etc.

Rules to Classify our Data:

If (Year < 2010) and (Country = Japan or German) and (Make = BMW or Audi or Ford or Honda or Citroen or Hyundai) THAN PASS = YES (0.5/0.0)

If (Year > 2010) and (Country = USA or UK or France or Czech Republic or Romania or (…) and (Make = Jaguar or Jeep or Chevrolet or Daewoo, Dacia or (…) THAN PASS = NO (0.4/0.0)

Several Outputs: If (Vehicle and Safety Equipment % >0) = Failed

or (Lighting and Electrical % >0) = Failed

or (…) THAN PASS = NO (0.9/0.0)

Ideally we would like to set probability at 40%, 50% and 90% level. The 40% for not passing test for some cars with lowest pass rate and with a year with lowest pass rate. Accordingly other probability rates have been set.

Of course it will be much more complicated because after the “Make” we would like to extract a model of car and also which test can still be failed. It is important information as fixing some items can be very expensive but other items can be very cheap.

Our goal is to find cars with passing ratio above 50% for all models and years of this car. So there is more than 50% that car will pass NCT test.

# Models:

## Naïve Bayes

Decision was made to use only few variables to classification rule to determine fail or pass. And then based on that determine what is typical test which this car fails.

Our important variables for classification: Vehicle Model, Vehicle Make, Year, Country, Total cars taken test and Pass rate and variables of test.

After spend some and reading about classification algorithms our choice was Naïve Bayes classifier.

Firstly because it is widely used and we wanted to see how it works and gets some experience about this algorithm. And secondly it seems that it is fast and reliable, and seems to be perfect for our project.

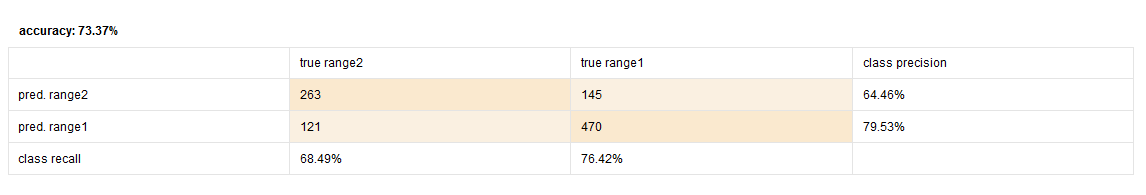
Naive Bayes it’s actually many algorithms based on Bayes Theorem. It allows predict an outcome, using probability.

We used 66% of data for learning purposes - 5000 rows.

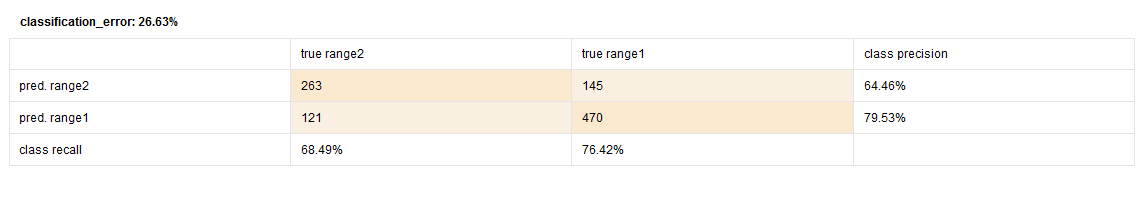
We divided our data into 2 ranges: Range 2 – pass more than 50% of this given car.

Range 1 – Pass less than 50% of this given car.

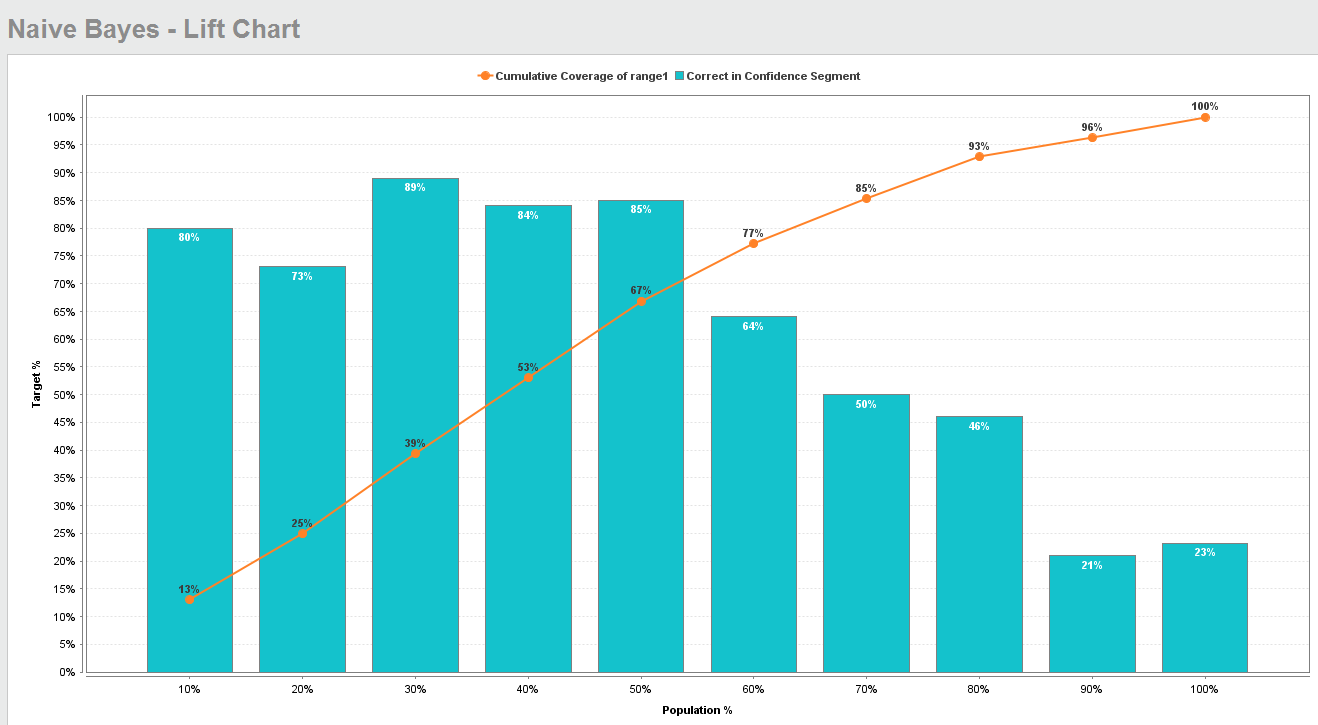
Naïve Bayes performance – ACCuracy 73.37%:

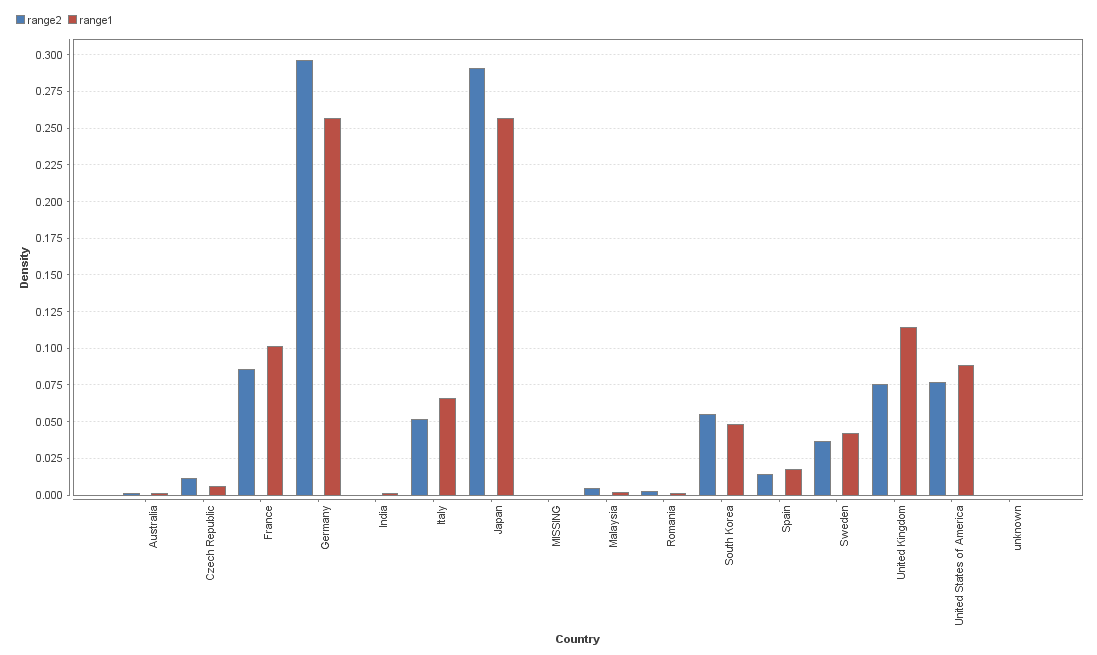


Naïve Bayes classification error 26.63%:



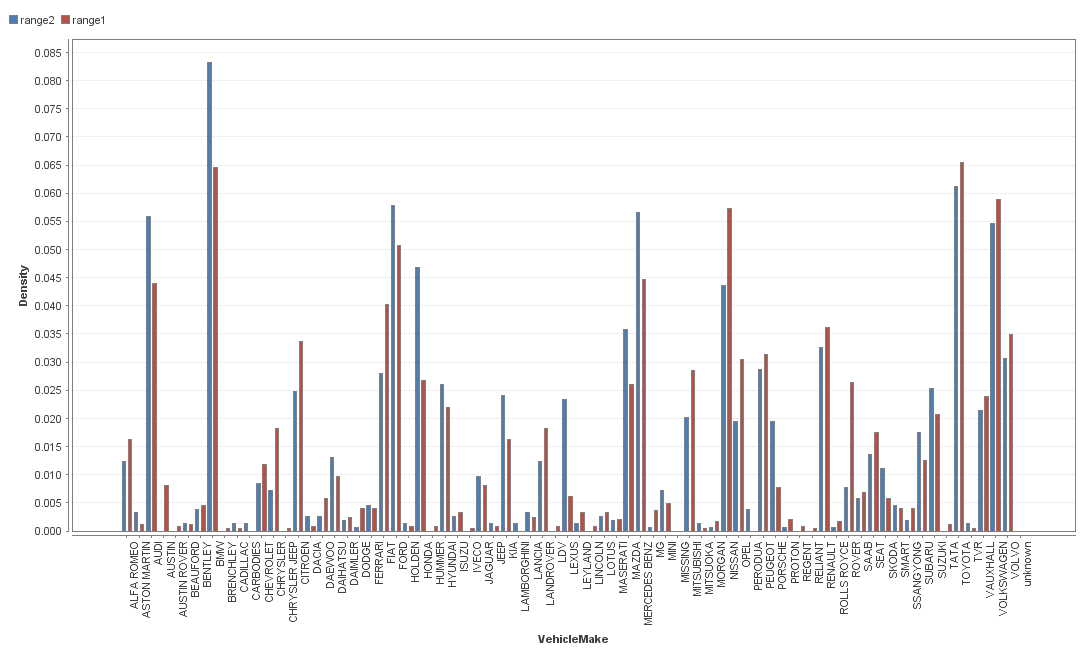
Naïve Bayes Lift Chart:



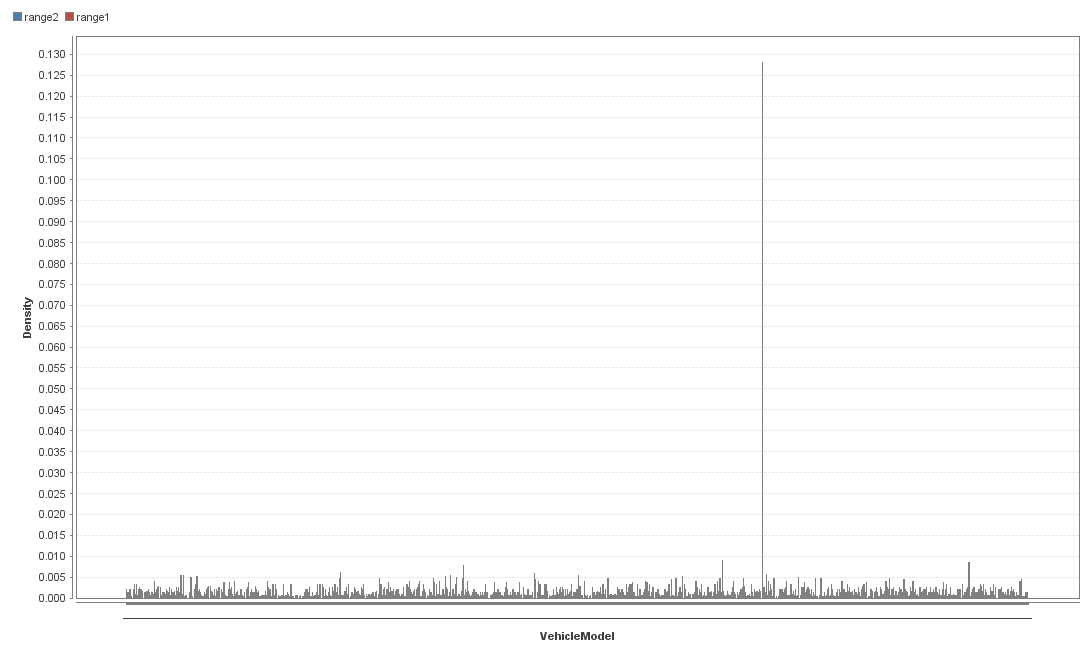
Model developed by RapidMiner:

Polynomial Attribute – Country

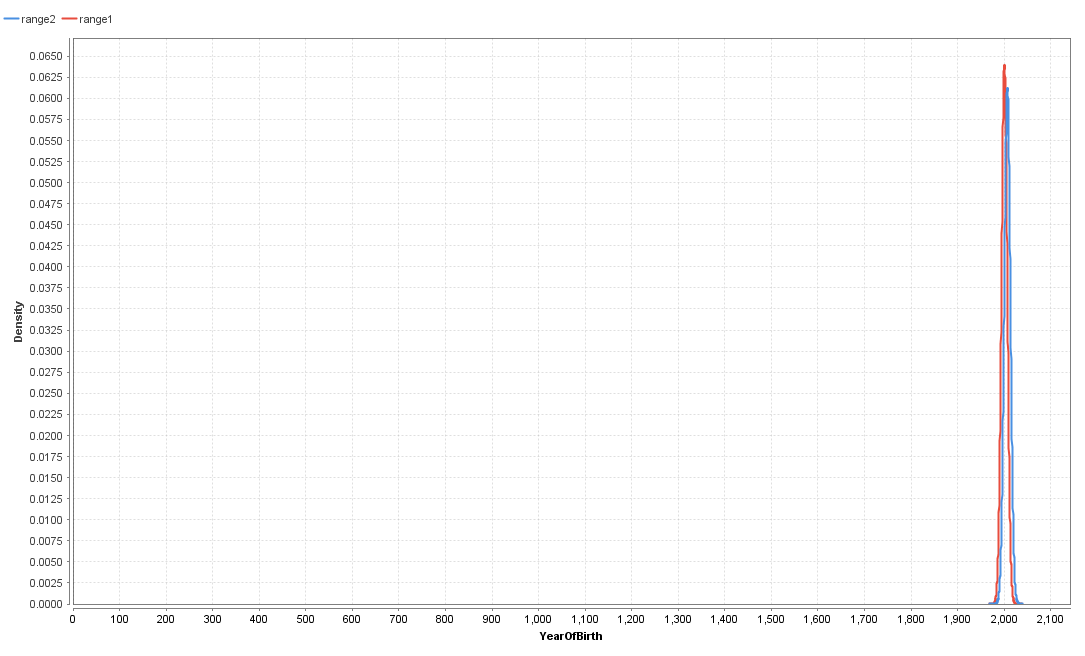
Polynomial Attribute – Vehicle Make



Polynomial Attribute – Vehicle Model



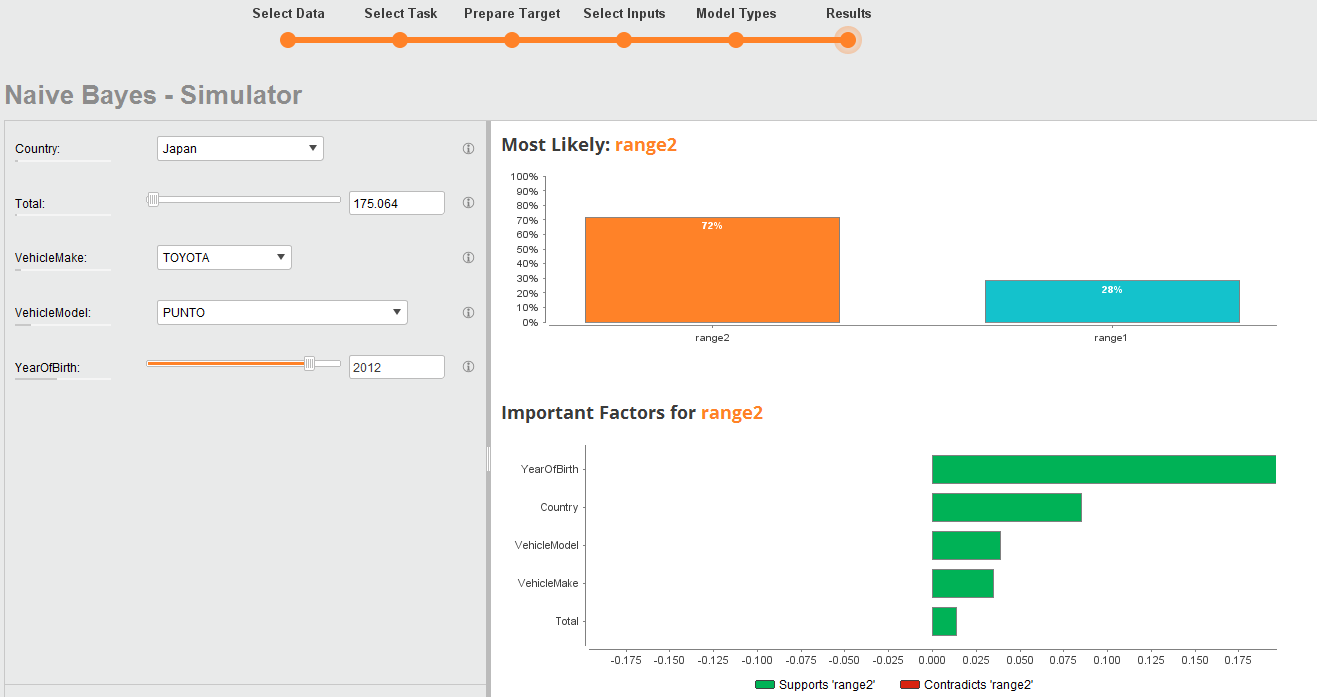
Attribute – Year



Simulator

When Creation of Model based on Naïve Bayes algorithm in Rapidminer Studio, we also were able to create simulator of make and Model and Range output.

Unfortunate we have discovered a big drawback of this classifier. Naïve Bayes assume all attributes to be independent. In simulator we could create prediction of pass rate of fleet of Toyota Punto of 2012. That’s a funny but not very helpful. Below you can see example.



In this particular case because of this drawback we find Naïve Bayes classifier not useful.

It is easy to determine PASS and not PASS result of NCT, but now we need to figure out another possible outcome. Reason for failing. Of course we could set up 14 similar models with different outcome – instead of PASS result, put as outcome the every single test. But that is very time consuming. It is better to create loop within algorithm and do it automatically.

Decision tree is much more helpful model in this problem as decision tree can have other outcomes than just 2 possibilities.

# Bibliography

1. <https://rapidminer.com/blog/naive-bayes-not-naive/>

1. <https://data.gov.ie/dataset/2016-make-model-year-failures-at-nct> [↑](#footnote-ref-1)
2. http://www.rsa.ie/Documents/VS\_Information\_Notes/Alternative\_Cars\_Fuels/FAQs%20on%20Vintage%20Vehicles.pdf [↑](#footnote-ref-2)