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

1. Introduction (including Problem statement, Goal).

Problem statement – given the car model and year, what is probability that it will fail NCT test and what kind of problems with car we can expect?

In Business goal – For example Taxi Company want to invest in new cars but one of factors for decision making will be NCT pass rate. We can help them to choose a correct car.

1. Data Description (Magda)
2. Data Preparation (NAN, OUTLIERS if any) KPI set up, ratios, percentages, binomial and nominal attributes. (If you have any suggestion to add column say in whatsup to make sure that we all know and add changes to github – although I think we have already calculated perfect ratios for all attributes of car test) (Magda)
3. Data mining model construction – define classification rules for models.

Car portfolio Segmentation – (suggested by Magda)

* 1. Cluster analysis
  2. Naïve Bayes classifier (Magda)

(How about use 50% data for training and then 30% for validation and 20% for testing?)

1. Data mining Solution (suggested by Magda)
   1. Logistic regression
   2. Decision Tree (maybe CART)
   3. Random Forest – improved version of decision tree
   4. Comparison of models and errors.
2. Conclusions ( including evaluation of findings - benefits and limitations of data mining models, recommendation of model to client)
3. References
4. Appendices

PS. Rapidminer jest naprawde swetnym urzadzeniem. Polecem:

<https://www.youtube.com/watch?v=U3FVLqV5Jzg> - Tutorial – Decision Tree in Rapidminer

<https://www.youtube.com/watch?v=0npgGPK0zR8> - Tutorial – Logistic regression in Rapidminer