# Alexey Antipovskiy

# Coursera Capstone project, 2020.

# Introduction

## A description of the problem and a discussion of the background.

We live in an environmental world. We can not imagine our life without machines but some of them can be quite dangerous, cause injuries or even human death when used improperly.

For example, according to WHO (World Health Organization) road accidents cause up to 1.4 million deaths worldwide annualy. Road accident is the 8th by frequency reason of death and the first one in 5-29 age period. People all over the world are working on the problem of minimisation of accident probability and severity because humans life is the most precious thing on Earth.

In our work we will try to analyze the collected data from car accidents statistics to find out

"How might we predict the probability of a road accident in actual conditions to warn the human about the upcoming danger before the road accident occurs?"

# Data

## A description of the data and how it will be used to solve the problem.

According to the business question we need to investigate Data, containing information about different types of road accidents, including information about seveity and different conditions that may have caused the accident. In this work we will explore the dataset, provided by Coursera and containing various information avout road accidents in Seattle& The dataset has 38 columns and 194 673 rows that provide multiple information concerning road accidents.We assume that we can pick relevant indicators to build a supervised model that will be able to predict the possible road accident with sufficient probability. So the general plan is:

* Inspect the dataset
* Fix missing data and type mismatch problems
* Investigate relations between parameeters and a target variable (Severity)
* Pick the most illustrative parameters
* Build a model based on picked parameters
* Check model quality
* Improve the model
* DEliver the final model with recomendations

# Dataset description

Initial dataset consists of 194673 rows and 38 columns. Let's check datatypes:

SEVERITYCODE int64

X float64

Y float64

OBJECTID int64

INCKEY int64

COLDETKEY int64

REPORTNO object

STATUS object

ADDRTYPE object

INTKEY float64

LOCATION object

EXCEPTRSNCODE object

EXCEPTRSNDESC object

SEVERITYCODE.1 int64

SEVERITYDESC object

COLLISIONTYPE object

PERSONCOUNT int64

PEDCOUNT int64

PEDCYLCOUNT int64

VEHCOUNT int64

INCDATE object

INCDTTM object

JUNCTIONTYPE object

SDOT\_COLCODE int64

SDOT\_COLDESC object

INATTENTIONIND object

UNDERINFL object

WEATHER object

ROADCOND object

LIGHTCOND object

PEDROWNOTGRNT object

SDOTCOLNUM float64

SPEEDING object

ST\_COLCODE object

ST\_COLDESC object

SEGLANEKEY int64

CROSSWALKKEY int64

HITPARKEDCAR object

dtype: object

We have selected these parameters to describe our model. The parameters have been chosen taking in mind the fact that they should be measurable during driving. They are:

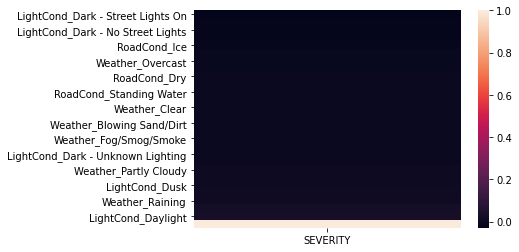
* Weather
* Road conditions
* Light conditions
* Speeding

and the dependent variable SEVERITY. Of course other parameters may play a very important role in our investigation but they can't be measured during driving (for example "Drugs influence" and the parameters of predicted incident). Let's construct the reduced datased excluding rows with Unknown or missing data.

After our operations dataset consists of 70510 rows and 5 columns. Let us perform one hot encoding as our variables mainly are categorical.

# Methodology

Let us examine the correlation of our variables with the dependent variable to choose appropriate model. We will use the Pearson correlation for this operation.



# Discussion section

Unfortunately, we have found out that none of the variables are in good corellation with Severity. That might mean that linear model approximation won't perform well. Let us examine top 5 and bottom 5 parameters that are in strongest correlation with SEVERITY.

|  | **SEVERITY** |
| --- | --- |
| **LightCond\_Dark - Street Lights On** | -0.029372 |
| **RoadCond\_Snow/Slush** | -0.021316 |
| **LightCond\_Dark - No Street Lights** | -0.019214 |
| **Weather\_Snowing** | -0.019056 |
| **RoadCond\_Ice** | -0.014229 |
| **RoadCond\_Wet** | 0.010186 |
| **Weather\_Raining** | 0.012265 |
| **SPEEDING** | 0.026915 |
| **LightCond\_Daylight** | 0.031597 |

As we can see the strongest correlation with speeding is only 3.2%. That means that non-linear approaches might result in better prediction accuracy.

# Statistical testing

Let us construct our models, defining Test and train datasets

Below the size of Train and Test arrays is provided:

Train arrays:(136408, 27) (136408,),

Test arrays (34102, 27) (34102,)

Let us build several models and examine their quality on Train and Test Data. Before we should Normalize Data.

## Linear Ridge Regression

Accuracy Ridge Linear Regression train:0.0033

Accuracy Ridge Linear Regression test:0.0035

As we have predicted linear model doesn't give good results because of low correlation.

## Logistic Regression

Accuracy Logistic Regression train:0.6722

Accuracy Logistic Regression test:0.6732

Logistic regression gives better results because of possible non-linear correlation between variables.

# Decision Trees

Accuracy Tree train:0.6723

Accuracy Tree test:0.6732

Decision trees give almost the same result as logistic regression - 67% accuracy

## K Nearest Neighbours

Accuracy Neigh train:0.6721

Accuracy Neigh test:0.6733

# Results section

As we have noticed there is no linear relationship between selected variables and a target variable. That is why we have achieved poor results using Linear Regression method. This problem was brilliantly solved using non linear method of Logistic Regression, Decision Trees. Both methods led to 67% accuracy on train and test Datasets. K Nearest neighbours resulted in approximately the same accurcy.

**Of course this result might be advanced but still this is already a good accuracy taking in mind that we deal with health and even the cost of a human life.**

# Conclusion section

We have finished examination of Accidents dataset and found out that predicting a car accident severity may be a complicated problem bacause the relationship between chosen variables and a target variable is not linear. A predictive model giving good results in 2/3 of cases was implemented using Logistics Regression and later proven by Decision Trees and K Nearest Neighbours.

We can use this output model in perspective automobiles to give signals to the driver after multiple variables analysis. And probably using this model will result in a rapid fall of deaths and injuries on the road (**2/3\*1.4 million = up to 1 million saved lives annually**)

**However a deeper investigation should take place because the initial dataset was a learning one and a lot of vital parameters such as Tire pressure, Drivers Experience, Car drive type (rear/front/4wd) were missing. Possibly using those missing arguments may result in better predictive models implementation.**