# Predicting the relationships between a car accident and weather conditions

## 1. Introduction

### 1.1 Background

Over the past few decades developing countries witnessed an excessive growth in population associated with large percentage of car ownership due to economic prosperity and reduction in car prices over time. This increased dramatically the traffic volumes on the street networks and caused - since the rate of increase in car ownership is way faster that the rate of developing new transportation infrastructure - serious traffic problems such as road accidents, traffic jams and excessive time delays. Investigating traffic accidents - as a multi-dimensional problem of engineering, social and legislative nature -is a research priority for most countries in order to reduce both accidents severity and frequency by investigating the three accidents stages: pre-accident, time of accident and post -accident stage. Many developing countries are in the process of developing complete digital databases for traffic accidents with all related attributes to be used in comprehensive analysis aiming to identify mainly the reasons behind traffic accidents and the best countermeasures to reduce the extent of this phenomenon [Sharaf A. Alkheder*; Reem Sabouni, Hany El Naggar; Abdul Rahim Sabouni. Driver and vehicle type parameters' contribution to traffic safety in UAE - https://www.scielo.br/scielo.php?script=sci_arttext&pid=S2238-10312013000200021 ].

### 1.2 Problem

Data that might contribute to determining player improvement might include his performance last season, his age, his draft status, his position, and metrics that describe what kind of player he is. This project aims to predict whether and how much a player will improve the next season based on these data.

### 1.3 Interest

It is obvious that information services will be in an accurate forecast of this dependence. Other investigators, car drivers, may also be helpful and interested.

## 2. Data acquisition and cleaning

### 2.1 Data sources

Most type of car accident, light, road, speed data can be found in dataset here. In this dataset we can find actual information about car accidents, weather conditions, addresses.

### 2.2 Data cleaning

Data downloaded or scraped from multiple sources were combined into one table. There were a lot of missing values (numeric, string), because of lack of record keeping. I decided to only use data about type of car accidents and weather conditions.

There are several problems with the dataset. First, some columns have empty value or NaN. I decided that we need to convert null values into integer. After fixing these problem, I checked for outliers in the data. I found there were some extreme outliers, mostly caused by some types of small sample size problem.

This data doesn't actual for our analysis because we focused more on parameters about weather and type of car accident.

### 2.3 Feature selection

After data cleaning, there were 1890 samples and 38 features in the data. Upon examining the meaning of each feature, it was clear that there was some redundancy in the features. These two features contained very similar information (about weather and precipitation), with the difference being. In our analyses we focused more about precipitations. In order to fix this, I decided to keep all features that were rates in nature, and drop their cumulative counterparts.

After discarding redundant features, I inspected the correlation of independent variables, and found several pairs that were highly correlated (Pearson correlation coefficient > 0.9). From these highly correlated features, only one was kept, others were dropped from the dataset.

# 3. Exploratory Data Analysis

## 3.1 Calculation of target variable

Revealing the dependence of weather conditions and the number of car accidents is the most urgent. As a target variable, I decided to calculate the difference of these parameters in the selected database. The calculation of the distribution of this dependence had a normal distribution centered around 0, with most values from -6 to 6. This suggests that the chosen metric for identifying the dependence of indicators was reasonable.

## 3.2 Relationship between car accidents and weather conditions

It is widely accepted that there is a better chance of having a car accident in bad weather conditions, and this is indeed borne out by our data. The median improvement in player performance declined with an increase in the number of crashes (z-test, p35, $p = 0.004$).

# 4. Predictive Modeling

In our work, we use a regression model that can be used to predict the reduction in the number of car accidents. Regression models can provide additional information about the degree of improvement.

### 4.1.1 Applying standard algorithms and their problems

I applied linear models (linear regression, Ridge regression), support vector machines (SVM), random forest, and gradient boost models to the dataset, using root mean squared error (RMSE) as the tuning and evaluation metric. The results all had the same problems. The predicted values had much narrow range than the actual values, and as a result, the prediction errors were larger as the actual values deviated further from zero. Having larger errors on those predictions was obviously not desirable.

### 4.1.2 Solution to the problems

My solution to this problem was to assign weights to samples based on the inverse of the abundances of target values. Using this method, all models predicted target values with similar range and distribution as the actual target values.

## 5. Methodology

In this project, we will focus our efforts on detecting the number of car accidents from the daily cycle, weather conditions and the type of accidents. At the first stage, we collected the necessary data: the number of car accidents from the daily cycle, weather conditions and the type of accidents. The second step in our analysis will be the calculation and research the percentage of each indicator in order to identify the leading positions for each item. In the third and final stage, we focused on creating a Linear Regression model of the type of car accident depending on weather conditions. Confirming results and analyze data.

## 6. Conclusion

In conclusion of our analysis, it can be noted that we have proven a direct relationship between car accidents and weather conditions. The greatest number of accidents occurs in rainy weather when snow and ice are observed. Also, the main type of accidents under these weather conditions is one parked-one moving entering at angle, from same direction. It can tell us that when weather conditions so bad like snow or wet, you have so high possibilities to have car accident.

## 7. Future directions

Models in this study mainly focused on individual features. These interactions data are obviously more difficult to extract and quantify, but if optimized, could bring significant improvements to the models.