**Capstone Project: Does Weather Affect Seattle Car Accidents?**

Liv Sullivan

**Introduction**

Cars are a staple of urban environments, and as a result traffic jams and accidents begin to occur as well. Nobody wants to be in a car accident, which is why we wear seatbelts, put down our phones, and make sure we are sober before a drive. People disregard these basic safety precautions, but those that choose to follow them should be able to assure themselves that they are as safe as can be when driving a car. However, there is another factor that affects car safety that man cannot control: the weather. When driving in certain weather conditions, you may find it harder to maintain control of the car and stay out of an accident.

**Business Problem**

The objective of this project is to predict the possibility and severity of a potential car accident based on weather and road conditions in the city of Seattle, WA. Using data science methodology and machine learning techniques, this project aims to provide solutions to the business question: In Seattle, WA, based on the weather and road conditions of the day, would you recommend avoiding driving that day?

**Target Audience**

This project is useful to anybody who drives a car in Seattle, WA and wants to remain safe while driving. We cannot control the weather, only our reactions to it, so those people will want to know if they should risk driving on a day with potentially dangerous conditions.

**Data**

**To solve the problem, we will need the following data:**

* List of historical car accidents in Seattle, WA. This will provide us with historical data of car accidents that we can use to make predictions about potential ones.
* Data regarding the weather and road conditions, the severity, and other details of the aforementioned car accidents. This will provide us the data necessary to provide an answer to the business question.

**Data Source**

The source used for this project can be found by [Clicking here](https://s3.us.cloud-object-storage.appdomain.cloud/cf-courses-data/CognitiveClass/DP0701EN/version-2/Data-Collisions.csv). The data provides information about all collisions provided by the Seattle Police Department and recorded by Traffic Records from 2004 to the present.

**Data Cleaning and Feature Selection**

To make the data more useful, I first rid the dataset of unnecessary categories until there was only SEVERITYCODE (the target variable), WEATHER, ROADCOND (road conditions), and LIGHTCOND (light conditions). Then, I removed all rows with missing attributes, and I did not lose valuable information by removing those rows because there were more than enough entries in the data set. I then used label encoding to turn object variables WEATHER, ROADCOND, and LIGHTCOND into categorical integer variables.

**Methodology**

**Exploratory Data Analysis**

I did not perform a great amount of exploratory data analysis, but I did do a perfunctory check of the value counts of the SEVERITYCODE (target variable) data. There were two values, severity codes of 1 and 2, and those that equaled 1 heavily outnumbered those that equaled 2. To balance the dataset, I used undersampling to make both results have 57052 occurrences.

**Machine Learning**

I used the following machine learning models on the dataset: K-Nearest Neighbor (KNN), Decision Tree, and Logistic Regression. I split the dataset into training and testing sets, with 80% used for training and 20% used for testing. I also normalized the dataset with the preprocessing Standard Scaler. KNN predicted the SEVERITYCODE of an accident based on the most similar data points within K distance. The decision tree model evaluated all the possible outcomes in order to fully analyze all the possibilities of a decision (all the possible outcomes of different weather, road, and light conditions). Logistic regression worked because there were only two values for the target variable (SEVERITYCODE), so the data was binary, and it estimated the probability of either result.

**Results**

I calculated the accuracy of the KNN model using the accuracy classification score, the jaccard similarity score, and the f1 score. The K that resulted in the most accuracy was K=5. The accuracy classification score 0.555 for the training set and 0.557 for the test set, the jaccard similarity score was 0.557, and the f1 score was 0.542.

I calculated the accuracy of the Decision Tree model using the accuracy classification score, the jaccard similarity score, and the f1 score. The accuracy classification score was 0.564, the jaccard similarity score was 0.564, and the f1 score was 0.540.

I calculated the accuracy of the Logistic Regression model using the jaccard similarity score, the f1 score, and the log loss. The jaccard similarity score was 0.535, the f1 score was 0.522, and the log loss was 0.681.

**Discussion**

Based on the results of all three models, the predictive capabilities of weather, road, and light conditions for the severity of a car accident are moderate. While sometimes certain weather will result in a car accident, other times it will not, and accidents occur in clear weather as well. Thus, weather conditions have a moderate impact on whether car travel could result in property damage (a severity code of 1) or injury (a severity code of 2). I recommend that a driver check the weather and make their own personal decision based off their own capabilities as a driver instead of relying on just the dataset evaluation results.

**Conclusion**

In this study, I analyzed the relationship between weather, road, and light conditions and the severity of car accidents in Seattle, WA. I identified that while there is sometimes correlation or causation, the results do not show high accuracy in predicting the severity of a car accident based on weather. I built classification models such as KNN, Decision Tree, and Logistic Regression models to identify this. In conclusion, weather cannot be used to accurately predict a car accident or its severity.