Add a waring feature to Navigation Apps

Cherry Cai

October 05, 2020

1.Introduction

1.1 Background

Given the dataset of accident records in Seattle, Washington district.A good way of utilizing it to benefit the local residents is to provide reference warning messages.The behavioral habit nowadays is to use navigation apps for direction. Hence navigation apps like Google Maps would be the best platforms to provide these messages. It is our target audience in this project.

1.2 Goal

The goal of this project is to predict the severity of road accidents at certain locations based on various conditions.When a user types in a destination and starts using the navigation function in the app, the app would prompt a warning if the place he is heading to or somewhere in the route is predicted to have a high chance of severe accident under the current conditions. The navigator could as well recommend a route that takes longer time but has lower risk to the user.

1.3 Benefits of this project

This would help the users reduce chances of getting fatal or severe injury. If a user receives the warning, he could cancel his plan or change to a safer route.On the other hand, It is a good opportunity for the tech companies to make use of big data resources to show humanistic care to the communities. Reducing chances of traffic accidents can in turn help saving costs of municipal services like police and ambulance.

2. Data

This project will use the sample "Data-Collisions.csv" file provided by the course to train the prediction model of traffic accidents' severity. The data contains accident records in Seattle of a date range from Jan 01, 2004 to May 02, 2020. There are 194673 events in the data set in total.

Based on the definition of our problem, the factors that will used in the prediction of severity of road accidents include:

- Geographical location
- Number of vehicles involved
- Weather factors (e.g., weather, road conditions, light conditions)
- Time of day
- Along with the severity level of the accident

These inputs are selected because they can be easily collected in real time at the moment an accident is reported.
As a result, the following attributes will be analyzed:

- The coordinates of the accident location X, Y
- VEHCOUNT which shows number of vehicles involved
- WEATHER, ROADCOND, LIGHTCOND
- INCDTTM
- SEVERITYCODE

Example of the data:

	X	Y	VEHCOUNT	WEATHER	ROADCOND	LIGHTCOND	INCDTTM	SEVERITYCODE
0	-122.323148	47.703140	2	Overcast	Wet	Daylight	3/27/2013 2:54:00 PM	2
1	-122.347294	47.647172	2	Raining	Wet	Dark - Street Lights On	12/20/2006 6:55:00 PM	1
2	-122.334540	47.607871	3	Overcast	Dry	Daylight	11/18/2004 10:20:00 AM	1
3	-122.334803	47.604803	3	Clear	Dry	Daylight	3/29/2013 9:26:00 AM	1
4	-122.306426	47.545739	2	Raining	Wet	Daylight	1/28/2004 8:04:00 AM	2

3. Methodology

3.1 Data Cleaning

There were lots of missing values because of lack of record keeping. After reading the data into pandas dataframe, all the features that stated above were selected to make a new dataframe which I mainly work on by giving a new name 'df_select'.

Firstly the record rows which lack any selected variable information (N/A) were dropped.This was because this new dataset would be used to train and test our model and all the selected features were considered to influence the final prediction. If they were kept, the columns which do not contain information would have to combine into other categories of the same feature which would cause bias to the final result.

Secondly the record rows which contain unhelpful information(Unknown or Other) were dropped. The reason is the same as above. Bias may occur. After these two steps, the data events dropped from 194673 to 166217.

Thirdly, weather conditions, road conditions and light conditions are various. Thus we need to combine these levels into less levels before we change them into numerical values. According to my natural cognition, weather and road conditions can be changed into boolean conditions and the light conditions can be changed into level numbers using LabelEncoder to represent different light levels.

Lastly, I was trying to parse the INCDTTM columns so that I can get the timestamp which represents time of the day info. When I ran the function to parse the time string, the system brought up an error stating that there were some rows of time string that did not contain the formatted time info. Hence, this kind of data rows were dropped.

3.2 Data Normalization

Normalize train/test set separately AFTER splitting.If normalize the whole dataset before splitting, the data that will be used as test set were already affected by the data in the training set when normalizing.

3.3 Model training

The problem will be solved using a logistic regression model since it is a classification model.

4. Results and Discussion

In the results above, we used Logistic Regression to attempt to add a warning function of Seattle Washington district in navigation apps.The Logistic Regression model we used has up to 66.8% accuracy of predicting traffic severity levels in Seattle.

During preprocessing the data, I deleted all the row data which don't have explicit information in the variables that were selected even though they noted 'Unknown' or 'Other'. The reason was because this kind of information would give no indication compared to other categories but would affect the final accuracy if I combine them into other categories which have apparent indication.

Thus due to strict data info requirements of the dataset, the original dataset dropped from 194673 rows to 143568 rows before it got trained. According to this, the apps should be able to access other APIs to guarantee that all required variables can be provided other than X/Y coordinates when using this model to provide predictions of severity of accidents.

5. Conclusion

Purpose of this project was to utilize the collision data records of Seattle to build a model which can predict the severity of road accidents in this area so that it can help the Seattle residents reduce chances of getting fatal or severe injury.

The prediction of this model would be used to give navigation users a warning if necessary. But the final decision of whether to go out or not after receiving a severity warning would still be made by individuals based on their personal requirements and needs.

To evaluate the model performance in real life, not only the prediction accuracy counts, additional factors should be taken into consideration.