Capstone Project – Predict Car Accident Severity

JENNY WONG

26 OCT, 2020

Introduction

- Life is the most valuable and irreplaceable thing in the world. I believe no one wants to have car accidents while travelling on road.
- To predict the severity of a car accident by using various data of historical car accidents like weather, date, road conditions, etc.
- To let people drive more carefully or even change the route or date of the journey

Data acquisition and cleaning

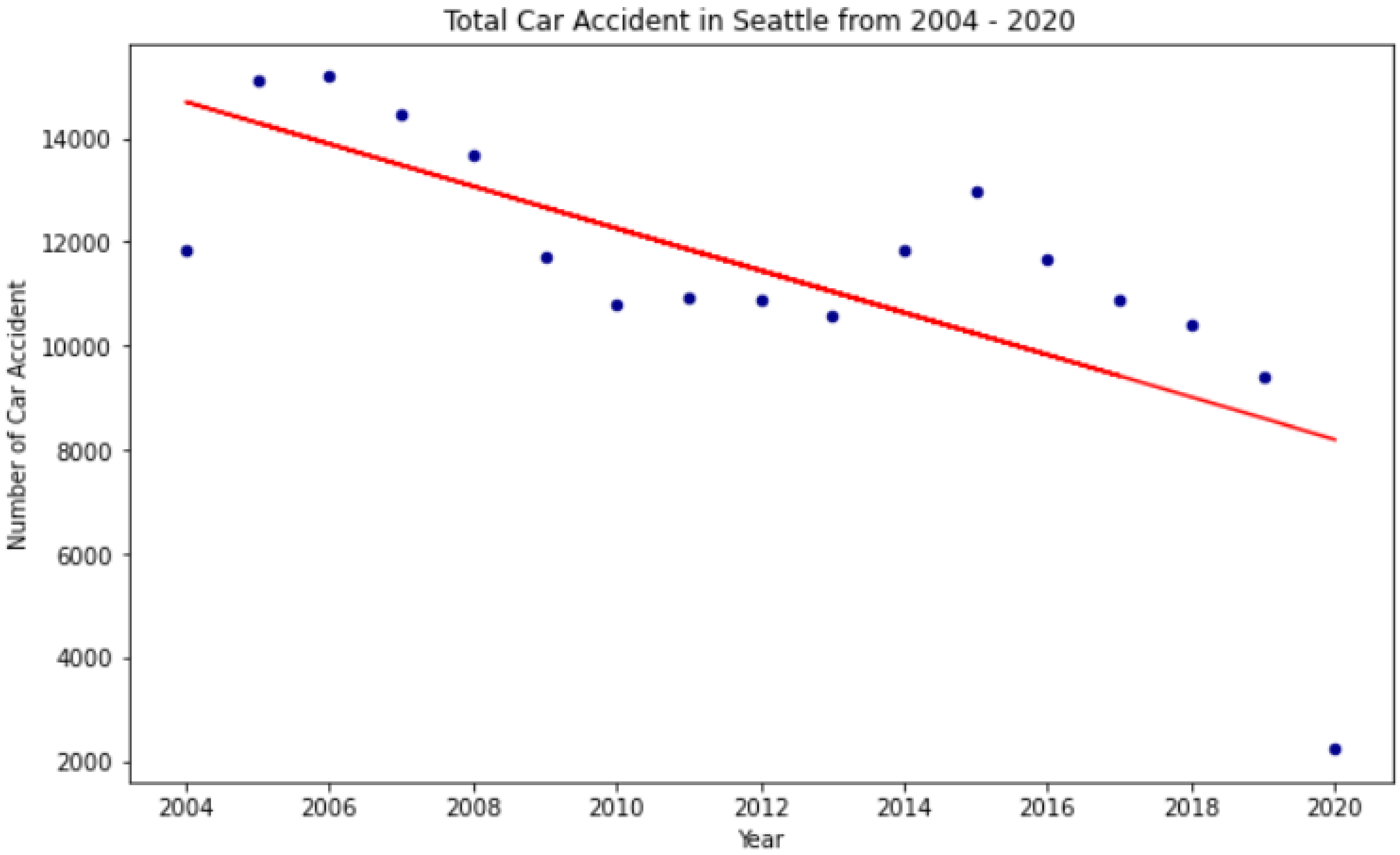
❑ Dataset is the traffic collisions in the City of Seattle (2004 – 2020).

❑ Provided by the Seattle Policy Department (SPD) and recorded by the Seattle Department of Transportation (SDOT) Traffic Management Division, Traffic Records Group. Download from here

❑ There are 194673 rows and 38 features in the raw dataset.

❑ Duplicated and highly similar features are dropped.

Total no. of car accident from 2004 to 2020



- ❑ The trend of number of car accident was dropping from 2005 to o 2013
- ❑ No. of car accident raised after 2013 and dropped again from 2014 to 2019
- ❑ Although data for 2020 is not a full set, can predict number of car accident 2020 keeps dropping,

Feature Selection

❑ Highly correlated features are used to compare the values with each of the others.

❑ Further analysis on the field JUNCTIONTYPE with ADDRTYPE

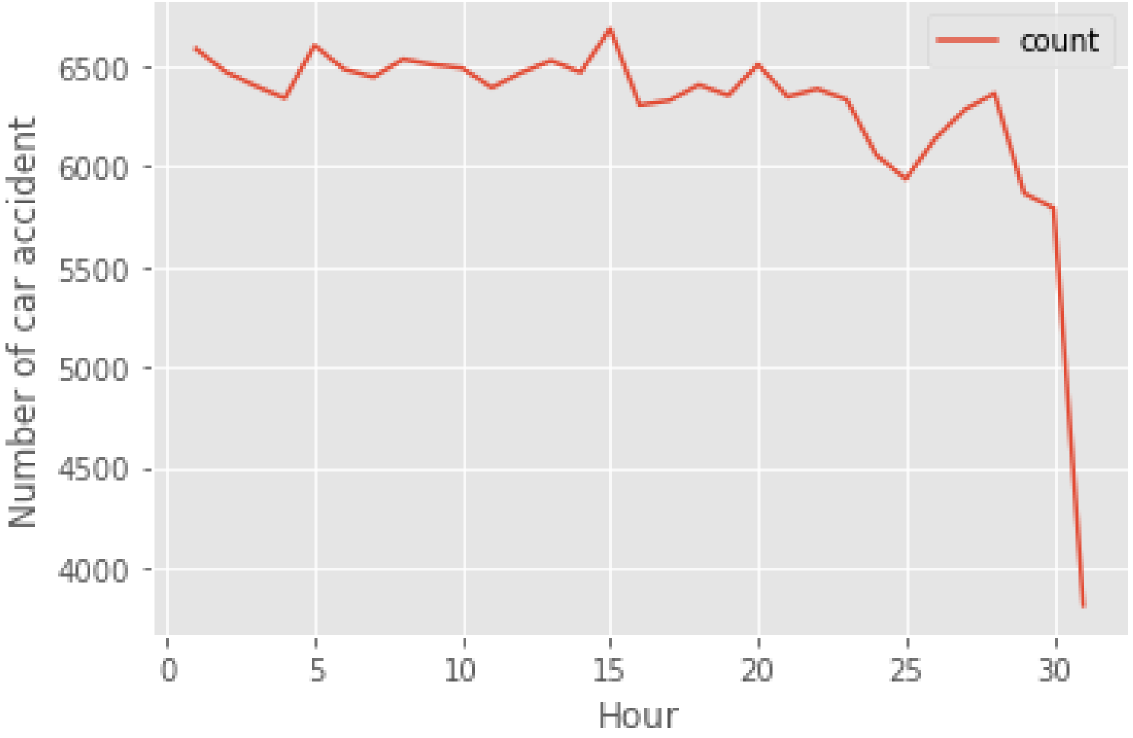
❑ Further analysis on the field WEATHER with ROADCOND

❑ Attribute from the raw dataset - ‘SEVERITYCODE', ‘INCDTTM’, 'JUNCTIONTYPE', ‘ADDRTYPE’, 'WEATHER', 'ROADCOND' and 'LIGHTCOND’

❑ Dropped all ‘Unknown’ value of the attributes 'WEATHER', 'ROADCOND' and 'LIGHTCOND’.

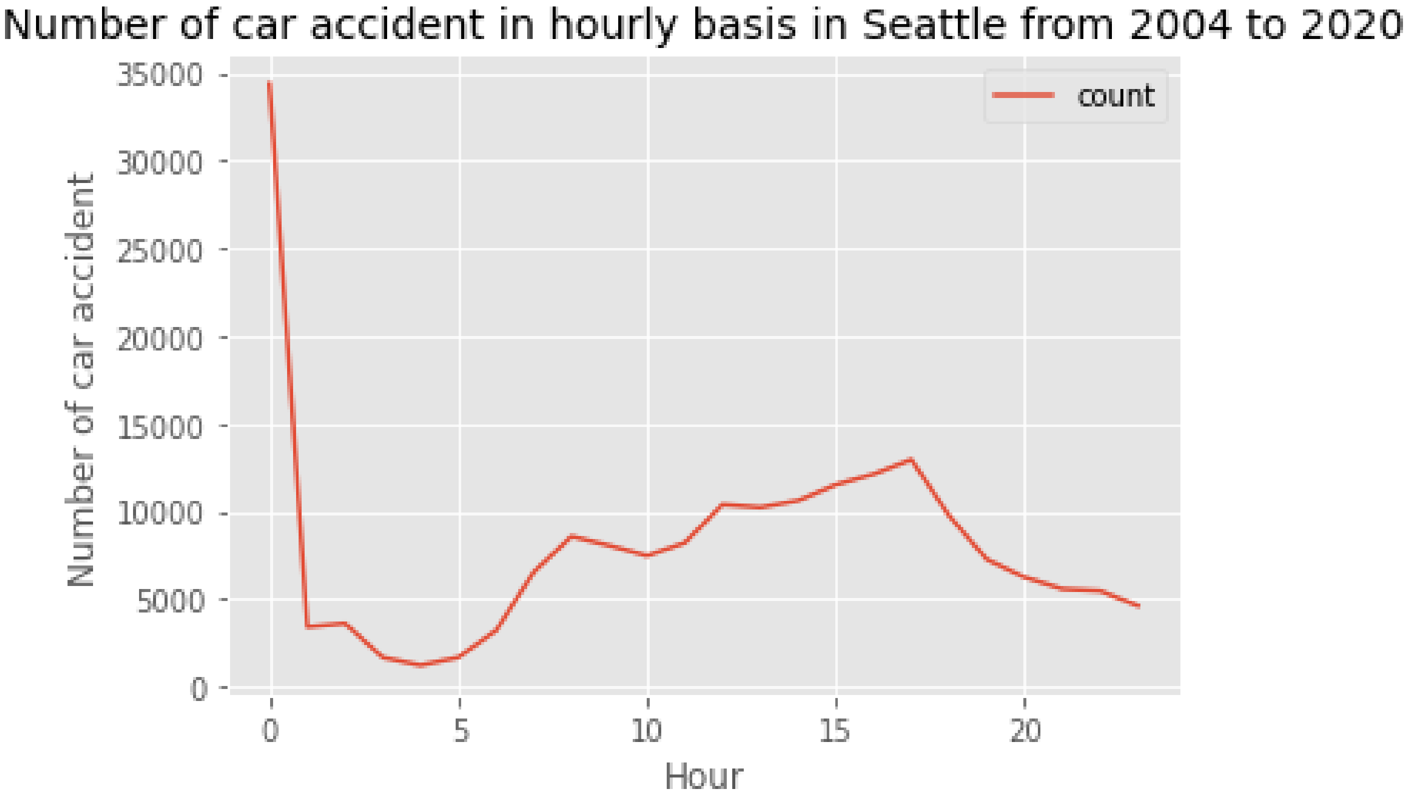
Analysis on INCDTTM (day)

Number of car accident in daily basis in Seattle from 2004 to 2020



- ❑ The number of car accident is quite steadily kept at around 6000 to 6500 per day
- ❑ Only the 31st of a month dropped almost half of the collisions because only 6 months with 31st of a year
- ❑ Not select this feature to predict the models

Analysis on INCDTTM (hour)



- ❑ The number of car accident extremely high at 00:00 to 01:00
- ❑ Another high is around15:00 to 17:00
- ❑ Take this feature to predict the models

Feature Selection Result

- ❑ Cleaned dataset contains 7 features.
- ❑ Selected 'JUNCTIONTYPE', ‘ADDRTYPE’, 'WEATHER', 'ROADCOND', 'LIGHTCOND' and ‘hour’ to predict ‘SEVERITYCODE’.
- ❑ Only 168659 rows of raw dataset are kept.

Predictive model selection

- ❑ Using supervised machine learning models to predict the severity of car accident.
- ❑ Classification algorithms :-
  - ❑ K Nearest Neighbor (KNN)
  - ❑ Decision Tree
  - ❑ Support Vector Machine
  - ❑ Logistic Regression
- ❑ Using 80% of dataset to train the models and the remaining 20% to test the models.

Models Evaluation

- Using 3 different evaluation metrics to calculate the accuracy of the models
  - Jaccard index
  - F1-score
  - Log Loss

Accuracy of predictive models

ALGORITHM	JACCARD INDEX	F1-SCORE	LOG LOSS
KNN	0.63	0.61	N.A.
Decision Tree	0.67	0.54	N.A.
SVM	0.67	0.54	N.A.
Logistic Regression	0.67	0.54	0.54

Conclusion

- ❑ Used 7 features to predict the severity of car accident.
- ❑ Used ~135000 records to train up the models and ~34000 to test the models.
- ❑ More accidents occurred when weather and road conditions are good (i.e. with enough light and not wet, a clear day).
- ❑ Interesting point: 12am to 1am had the most car accident happened; peak hour like off-duty period is the second highest.
- ❑ Should keep pay attention especially the weather and road conditions are good.