Cousera IBM Data Science Capstone Project

Car Accident Severity

Contents

- Introduction
- Data
  - Feature Selection
- Methodology
  - Exploratory Data Analysis
  - Feature Encoding
  - Feature Correlation Analysis
  - Model
- Results
- Summary
- Recommendations

Introduction

- Stakeholders:
  - Seattle Department of Transportation
  - Seattle Police Department - Traffic Enforcement Section
  - Car Drivers in Seattle
- Car accident is serious problem in Seattle. During the first six months of 2019, 101 people were seriously injured or killed in 98 collisions on Seattle streets, which is the highest number of crashes in the first half of a year since 2010, according to the data provided by the Seattle Department of Transportation (SDOT). The accident victims and their families, the insurance conpanies, health care personal and even the ordinary people, are affected by traffic accidents in many ways. Therefore, predicting the possibility and severity of a car accident based on the weather, road conditions and other factors is important. For drivers, they would drive more carefully or even change their travel if they are able to. For the police and traffic departments, they can put up some warning signs when there is a high traffic accident risk.

Data

	SEVERITYCODE	X	Y	OBJECTID	INCKEY	COLDETKEY	REPORTNO	STATUS	ADDRTYPE	INTKEY	LOCATION	EXCEPTRSNCODE	EXCEPTRSNDESC	SEVERITYCODE.1	SEVERITYDESC	COLLISIONTYPE
0	2	-122.323148	47.703140	1	1307	1307	3502005	Matched	Intersection	37475.0	5TH AVE NE AND NE 103RD ST		NaN	2	Injury Collision	Angles
1	1	-122.347294	47.647172	2	52200	52200	2607959	Matched	Block	NaN	AURORA BR BETWEEN RAYE ST AND BRIDGE WAY N	NaN	NaN	1	Property Damage Only Collision	Sideswipe
2	1	-122.334540	47.607871	3	26700	26700	1482393	Matched	Block	NaN	4TH AVE BETWEEN SENECA ST AND UNIVERSITY ST	NaN	NaN	1	Property Damage Only Collision	Parked Car
3	1	-122.334803	47.604803	4	1144	1144	3503937	Matched	Block	NaN	2ND AVE BETWEEN MARION ST AND MADISON ST		NaN	1	Property Damage Only Collision	Other
4	2	-122.306426	47.545739	5	17700	17700	1807429	Matched	Intersection	34387.0	SWIFT AVE S AND SWIFT AV OFF RP	NaN	NaN	2	Injury Collision	Angles

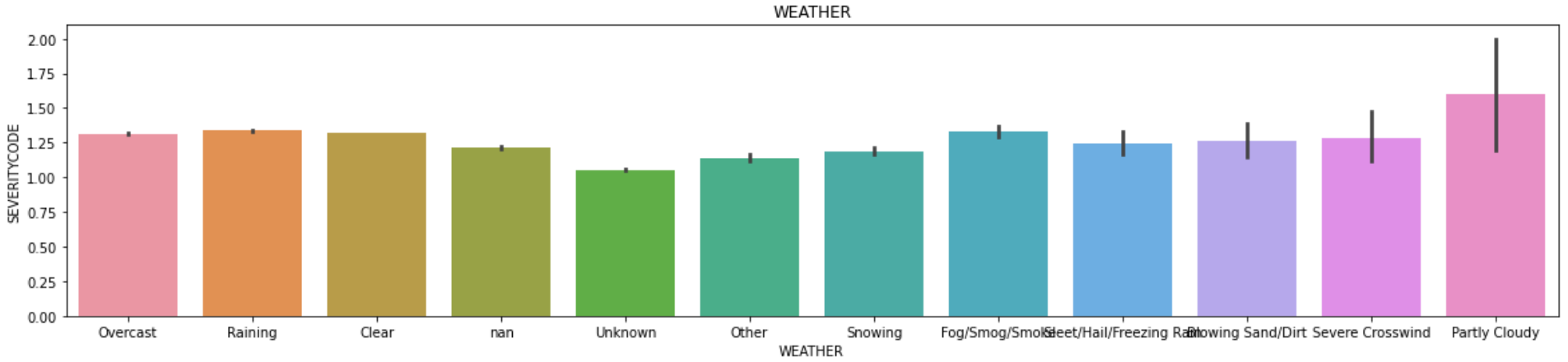
- The dataset used in this analysis is provided by SDOT Traffic Management Division, Traffic Records Group. It includes all types of collisions at the intersection or mid-block of a segment recorded by the Traffic Records from 2004 to Present.
- The dataset contains 194,673 records with 38 attributes: 16 numerical and 22 categorical (see the figures below for the detailed descriptions). Among these 38 features, 6 of them have a large portion of null value, which are INTKEY, EXCEPTRSNCODE, EXCEPTRSNDESC, INATTENTIONIND, PEDROWNOTGRNT, SPEEDING. We will perform imputation on these feature during the feature engineering step. The target for this prediction task is the "Accident Severity", which are represented by three attributes: SEVERITYCODE, SEVERITYCODE.1, SEVERITYDESC, containing two sets of values: 1, 1, Property Damage Only Collision and 2, 2, Injury Collision. The target is unevenly distributed, with ~70% Property Damage Collisions Only and ~30% Injury Collisions. Therefore, target weights will be applied in the modeling step.

Data: Feature Selection

- Target: Choose the 'SEVERITYCODE' (A code that corresponds to the severity of the collision: 1-Property Damage; 2-Injury), or 'SEVERITYCODE.1', 'SEVERITYDESC', as the target.
- Features: Choose features related to the weather, road and drivers' condition.
  - 'INATTENTIONIND': Whether or not collision was due to inattention.
  - 'UNDERINFL': Whether or not a driver involved was under the influence of drugs or alcohol.
  - 'WEATHER': A description of the weather conditions during the time of the collision.
  - 'ROADCOND': The condition of the road during the collision.
  - 'LIGHTCOND': The light conditions during the collision.
  - 'SPEEDING': Whether or not speeding was a factor in the collision.

Methodology

- Exploratory Data Analysis
  - The target is unevenly distributed, with ~70% Property Damage (1) and ~30% Injury (2).
  - Inattention ('INATTENTIONIND'='Y) will increase the collision severity.
  - Drivers under the influence of drugs or alcohol ('UNDERINFL'='Y' or '1') will increase the collision severity.
  - Speeding ('SPEEDING'='Y') will increase the collision severity.
  - Bad weather ('WEATHER'), road ('ROADCOND') and light ('LIGHTCOND') conditions will also slightly increase the collision severity.



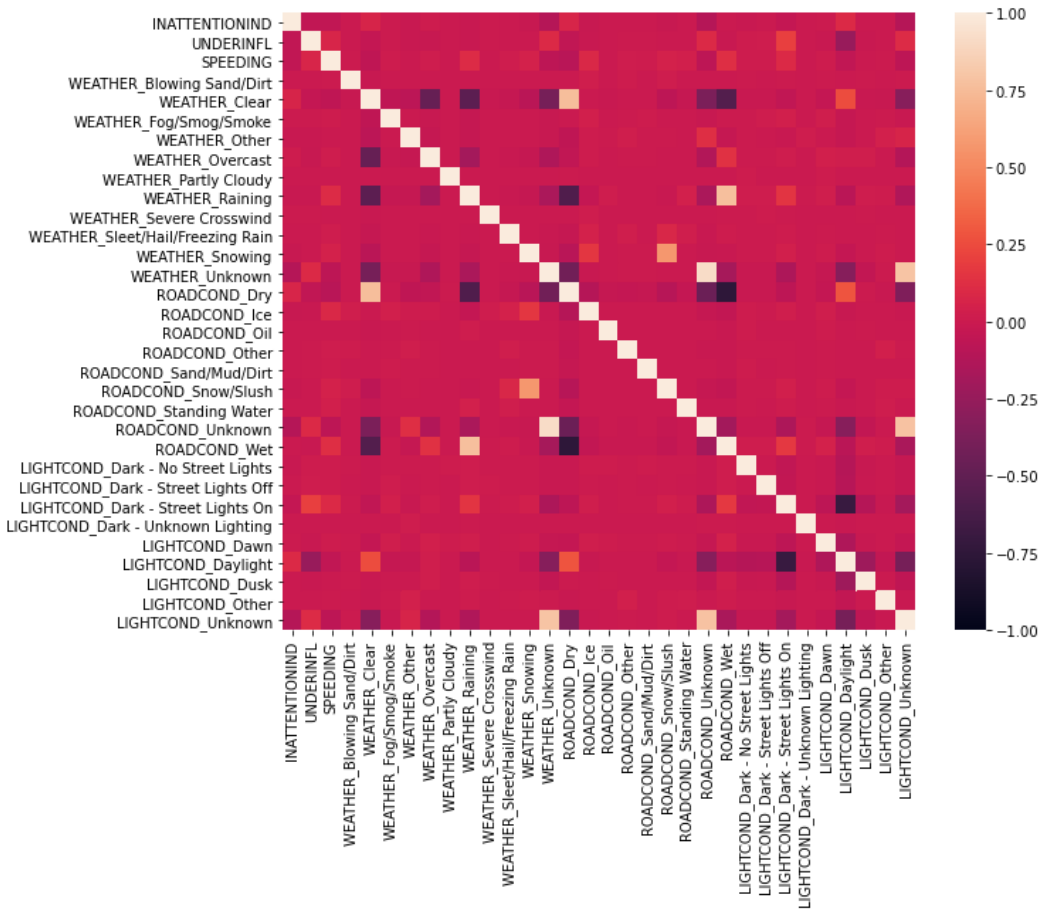
Methodology

- Feature Encoding
  - Use label encoding and one hot encoding
  - Output: 194673 samples with 32 features

	INATTENTIONIND	UNDERINFL	SPEEDING	SEVERITYCODE	WEATHER_Blowing Sand/Dirt	WEATHER_Clear	WEATHER_Fog/Smog/Smoke	WEATHER_Other	WEATHER_Overcast	WEATHER_Partly Cloudy	WEATHER_Raining	WEATHER_Severe Crosswinc
0	0	0	0	1	0	0	0	0	1	0	0	0
1	0	0	0	0	0	0	0	0	0	0	1	0
2	0	0	0	0	0	0	0	0	1	0	0	0
3	0	0	0	0	0	1	0	0	0	0	0	0
4	0	0	0	1	0	0	0	0	0	0	1	0

Methodology

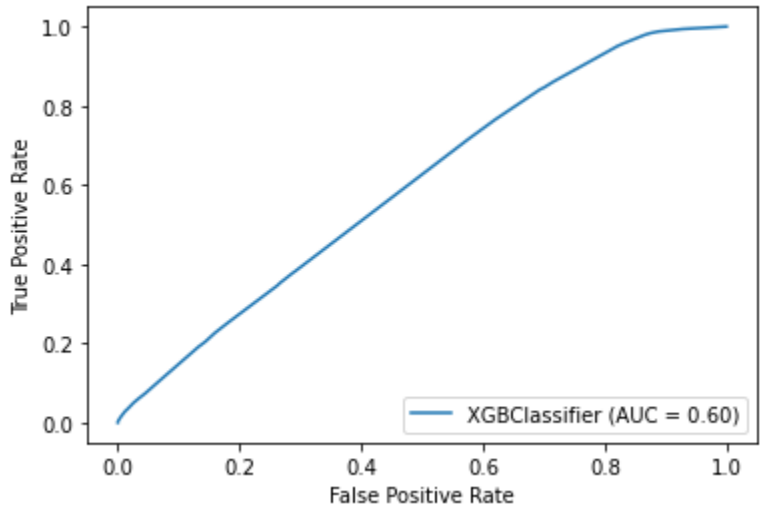
- Feature Correlation Analysis
  - High correlation observed between weather, road and light conditions, which is reasonable since the road and light conditions are both related to the weather conditon.



Methodology

- Model
  - Logistic Regression: simple baseline model
  - Decision Tree: simple tree based model
  - Random Forest: advanced tree based model using the bagging method
  - XGBoost: advanced tree based model using the boosting method

- Feature Preprocessing
  - feature standardization
  - oversampling
  - train & test splitting



Results

- Metrics: Precision, Recall ,F1-score, Accuracy, ROC-AUC, logloss
- Overall, the 4 models performs similarly. But the models metrics are not very good.
- The Random Forest and XGBoost perform slightly better than the Logistic Regression and Decision Tree.

	Precision	Recall	F1-score	Accuracy	ROC-AUC	logloss
Logistic Regression	0.549843	0.789464	0.648218	0.571565	0.593586	0.668194
DecisionTree	0.547677	0.809246	0.653251	0.570447	0.593542	0.663569
Random Forest	0.547307	0.855918	0.667676	0.573982	0.597957	0.663035
XGBoost	0.548157	0.839616	0.663281	0.573763	0.598076	0.661718

Summary

- We have performed Exploratory Data Analysis, Feature Engineering and Modeling for the project. Use 6 features related to the weather, road and driver's conditions to predict the severity of an accident.
- We examined 6 metrics: Precision, Recall ,F1-score, Accuracy, ROC-AUC, logloss, on 4 models: Logistic Regression, Decision Tree, Random Forest and XGBoost, where the XGboost model performs the best, with a Precision score of 0.548157, a Recall score of 0.839616, a F1 score of 0.663281, an accuracy score of 0.573763, a ROC-AUC score of 0.598076, and a logloss score of 0.661718.
- The overall performance of all the models are not very good. For improvement, we can add more features from the original dataset to our models.

Recommendations

- To the transportation department:
  - Improve the road and light conditions for those roads where most severe accidents take place.
  - Put up signs to remind driver to stay conscious, and slow down when weather condition is not optimal.
  - Send our more police officers to monitor the speeding issue.

- To the car drivers:
  - Pay attention to the change of road and weather condition.
  - Always be conscious and focused, and obey the traffic rules.