Coursera

Applied Data Science Capstone Project

Accident severity prediction in Seattle

Manlai Amarsaikhan

September 1, 2020

Today's talk

- Introduction
- Data
- Exploratory data analysis
- Classification strategy
- Results
- Discussion

Introduction

- Every commuter faces a non-negative probability of getting in a car accident everyday
- Car accidents are costly
- People try to mitigate the possible costs by buying insurance
- Best case scenario: Know the probability of getting in car accidents
- Next best case: Know the severity of getting in car accidents
  - The topic of this project

Data

- Data set from Seattle Police Department
- Records of all collisions in Seattle between 2004 and 2019
- 194,673 samples with 37 features and 1 label

Data cleaning

- Dropped duplicate samples
- Dealt with missing values as follows:
  - Corrected existing values that should have been NULL to NULL
  - Corrected NULL values that should have been 0 to 0
  - After above process, dropped all samples with missing values
- Duplicate or similar features were dropped
- Features whose information was contained in other features were also dropped

Data cleaning

In the end, there were 17 features and 1 label:

- Label: SEVERITYCODE
- Categorical variables: ADDRTYPE, COLLISIONTYPE, JUNCTIONTYPE, WEATHER, ROADCOND, LIGHTCOND
- Binary variables: INATTENTIONIND, UNDERINFL, PEDROWNOTGRNT, SPEEDING
- Continuous variable: X, Y, PERSONCOUNT, PEDCOUNT, PEDCYLCOUNT, VEHCOUNT, INCDTTM

From date-time variable INCDTTM, further 3 categorical values were derived: MONTH, DAYOFWEEK, HOUROFDAY

Exploratory data analysis

SEVERITYCODE vs ADDRTYPE

- $\chi^2 = 7546.62$, degrees of freedom$= 2$, $p = 0.00$

Table: Observed frequencies of SEVERITYCODE and ADDRTYPE

	ADDRTYPE		
SEVERITYCODE	**Alley**	**Block**	**Intersection**
1	669	96829	37251
2	82	30094	27819

Table: Expected frequencies of SEVERITYCODE and ADDRTYPE

	ADDRTYPE		
SEVERITYCODE	**Alley**	**Block**	**Intersection**
1	525.03	88732.97	45491
2	225.97	38190.03	19579

Exploratory data analysis

SEVERITYCODE vs COLLISIONTYPE

- $\chi^2 = 41075.56$, degrees of freedom$= 9$, $p = 0.00$

Table: Observed frequencies of SEVERITYCODE and COLLISIONTYPE

SEVERITYCODE	parked car, right turn, sideswipe or other	Angles, cycles, head on, left turn, pedestrian, or rear ended
1	81365	55119
2	11889	46297

Exploratory data analysis

SEVERITYCODE vs COLLISIONTYPE

Table: Expected frequencies of SEVERITYCODE and COLLISIONTYPE

SEVERITYCODE	**parked car, right turn, sideswipe or other**	**Angles, cycles, head on, left turn, pedestrian, or rear ended**
1	65380.79	71103.21
2	27873.21	30312.79

Exploratory data analysis

SEVERITYCODE vs ROADCOND

- $\chi^2 = 185.73$, degrees of freedom$= 7$, $p = 0.00$

Table: Observed frequencies of SEVERITYCODE and ROADCOND

ROADCOND	**Dry**	**Ice**	**Oil**	**Other**	**Sand** Mud Dirt	**Snow Slush**	**Standing Water**	**Wet**
1	84446	936	40	89	52	837	85	31719
2	40063	273	24	43	23	167	30	15754

Table: Expected frequencies of SEVERITYCODE and ROADCOND

ROADCOND	**Dry**	**Ice**	**Oil**	**Other**	**Sand** Mud Dirt	**Snow Slush**	**Standing Water**	**Wet**
1	84301.62	818.58	43.33	89.37	50.78	679.78	77.86	32142.66
2	40207.38	390.42	20.67	42.63	24.22	324.22	37.14	15330.34

Exploratory data analysis

SEVERITYCODE vs LIGHTCOND

- $\chi^2 = 284.07$, degrees of freedom$= 7$, $p = 0.0$

Table: Observed frequencies of SEVERITYCODE and LIGHTCOND

LIGHTCOND	Dark No Street Lights	Dark Street Lights Off	Dark Street Lights On	Dark Unknown Lighting	Dawn	Daylight	Dusk	Other
1	1203	883	34032	7	1678	77593	3958	183
2	334	316	14475	4	824	38542	1944	52

Table: Expected frequencies of SEVERITYCODE and LIGHTCOND

LIGHTCOND	Dark No Street Lights	Dark Street Lights Off	Dark Street Lights On	Dark Unknown Lighting	Dawn	Daylight	Dusk	Other
1	1043.75	814.22	32940.11	7.47	1699.06	78864.89	4007.93	159.58
2	493.25	384.78	15566.89	3.53	802.94	37270.11	1894.07	75.42

Classification strategy

- Created dummy variables from the categorical variables.
  - 54 features
- The sample split into sample set and validation set (with 4:1 ratio).
- Sample set further split into train set and test set (with 3:1 ratio)
- Oversampled using SMOTE method to deal with unbalanced data
- Accuracy and ROC AUC scores were chosen as metrics

Results

Table: Classification results

	Dummy	**LogisticRegression**	**GradientBoost**	**RandomForest**
accuracy	0.668	0.728	0.735	0.732
ROC AUC	0.5	0.753	0.766	0.764

Discussion

- All trained models perform better than the dummy (that predicts the majority label)
- Gradient Boosting Classifier performs the best
- Interaction terms are possibly needed to improve upon the current accuracies obtained
  - e.g. Rain during rush hour, speeding on certain locations etc.