Predicting collision vehicles in Seattle

Introduction

Business Problem¶

A system to reduce accidents will be favoured by car users, insurance companies, road maintenance, cities/municipalities/governments and, if relevant, habitants in the area. It would reduce the costs caused by accidents, as well as improve the traffic by reducing congestion caused by accidents. Predict the possibility of getting to a car accident and its severity given the current driver and driving conditions in order to reduce damages in a real-life scenario.

Data

Data Source:

- Data is already provided through a CSV file called Data-Collisions.csv which contains 194673 rows.

Feature Selection:

- The next step is to remove irrelevant columns: OBJECTID','INCKEY','COLDETKEY','INTKEY','SEVERITYCODE.1','SDOT_COLCODE','SDOTCOLNUM','SEGLANEKEY','CROSSWALKKEY','REPORTNO','STATUS','ADDRTYPE','LOCATION','EXCEPTRSNCODE','EXCEPTRSNDESC','INATTENTIONIND','UNDERINFL','PEDROWNOTGRNT','SPEEDING','ST_COLCODE','ST_COLDESC','HITPARKEDCAR','COLLISIONTYPE','INCDTTM','JUNCTIONTYPE','SDOT_COLDESC'.
- Then we have to identify the relevant columns by looking for trends and patterns in this case: WEATHER, ROADCOND, LIGHTCOND, PERSONCOUNT, PEDCOUNT, PEDCYLCOUNT, VEHCOUNT, INCDATE.

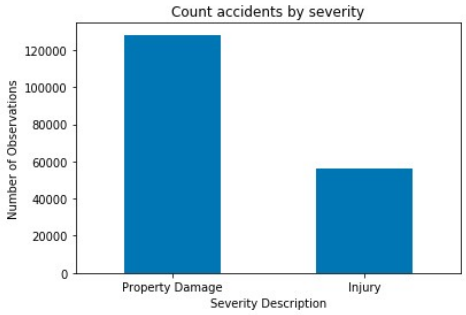
Data Cleaning:

- Dealing with missing data by removing the rows.
- Balancing in order to dealing with inbalanced data.

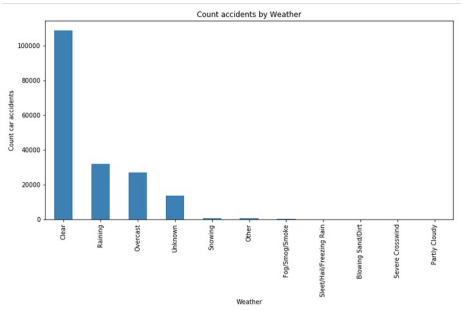
Methodology

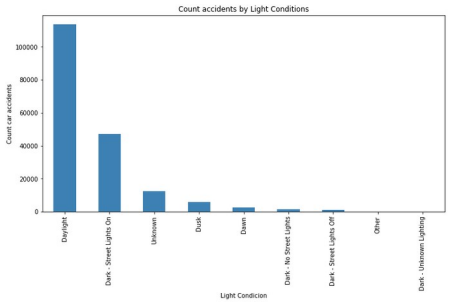
Exploratory analysis

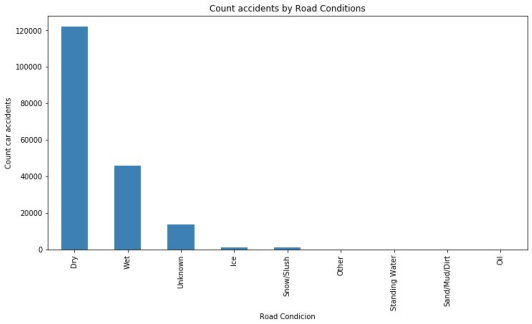
Show count accidents by severity.



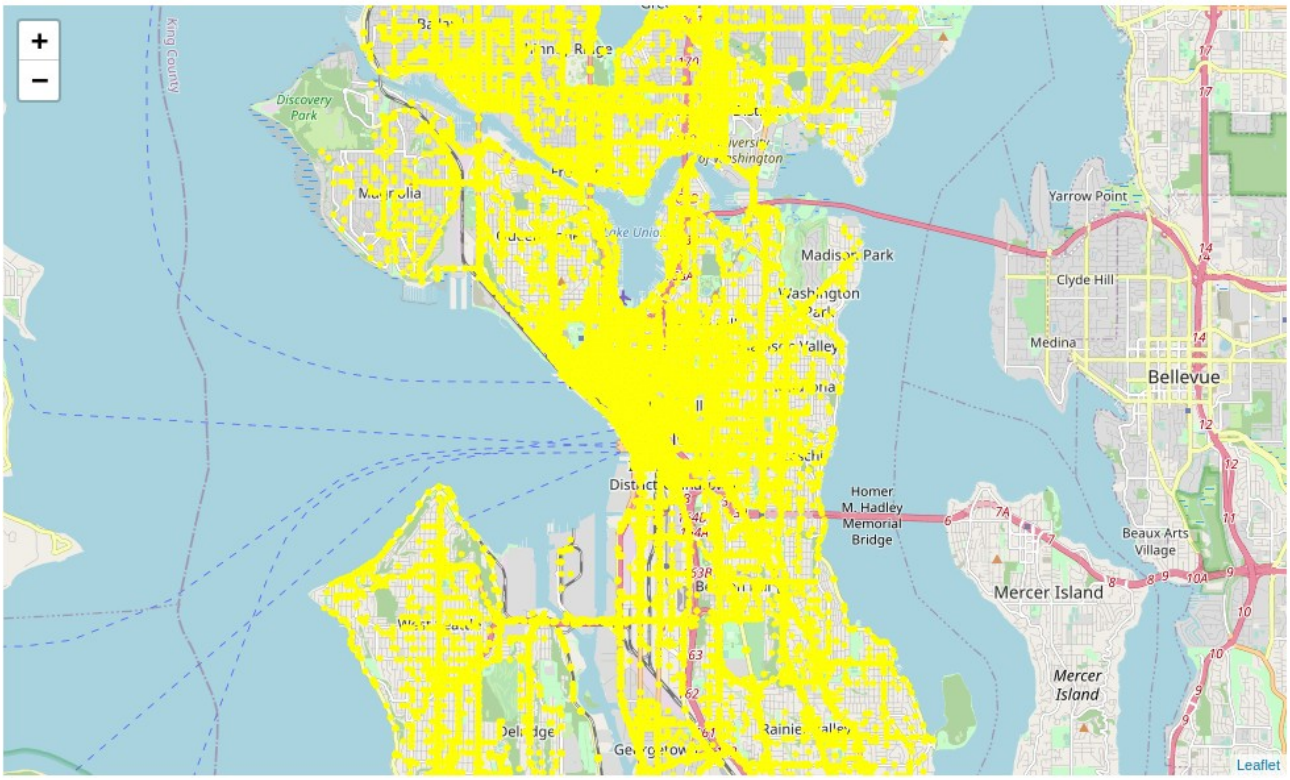
The majority accidents happened with clear weather, in a dry road and daylight.





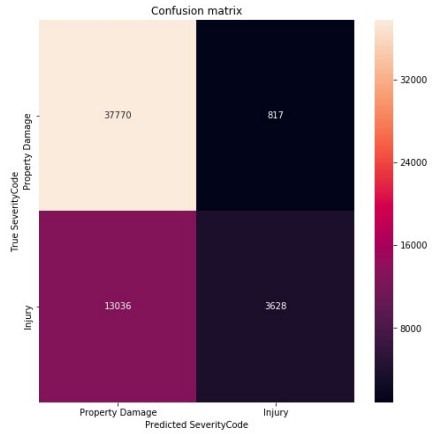


Show map with accident severitycode = 2 (physical injuries). They are spread over entire city.



Results

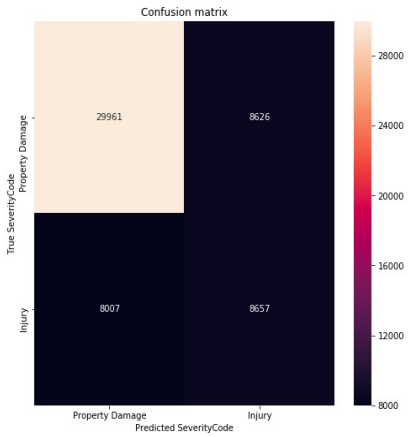
Logistic Regression with imbalanced data.



	precision	recall	f1-score	support
1	0.74	0.98	0.85	38587
2	0.82	0.22	0.34	16664
accuracy			0.75	55251
macro avg	0.78	0.60	0.59	55251
weighted avg	0.77	0.75	0.69	55251

Log Loss: 24.122253331459294

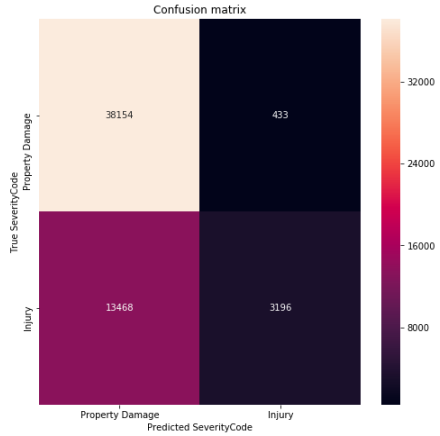
Logistic Regression with balanced data.



	precision	recall	f1-score	support
1	0.79	0.78	0.78	38587
2	0.50	0.52	0.51	16664
accuracy			0.70	55251
macro avg	0.65	0.65	0.65	55251
weighted avg	0.70	0.70	0.70	55251

Log Loss: 24.122253331459294

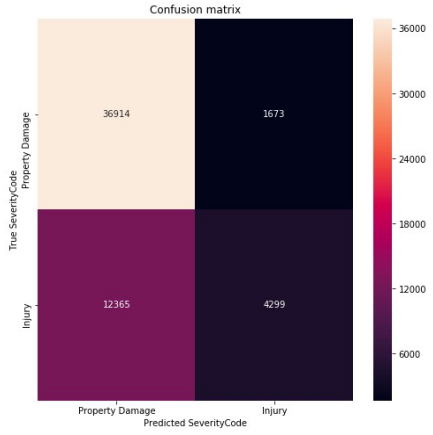
Decision tree with imbalanced data



	precision	recall	f1-score	support
1	0.74	0.99	0.85	38587
2	0.88	0.19	0.31	16664
accuracy			0.75	55251
macro avg	0.81	0.59	0.58	55251
weighted avg	0.78	0.75	0.69	55251

Log Loss: 24.122253331459294

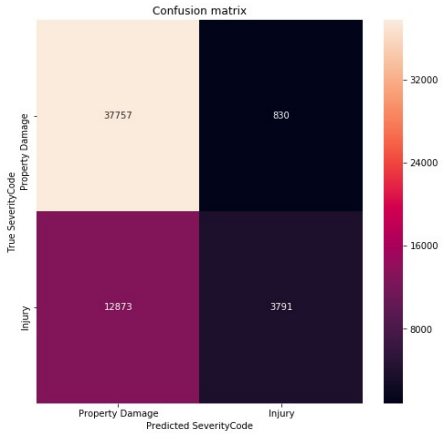
KNN & the best k (6) – imbalanced data



	precision	recall	f1-score	support
1	0.75	0.96	0.84	38587
2	0.72	0.26	0.38	16664
accuracy			0.75	55251
macro avg	0.73	0.61	0.61	55251
weighted avg	0.74	0.75	0.70	55251

Log Loss: 24.122253331459294

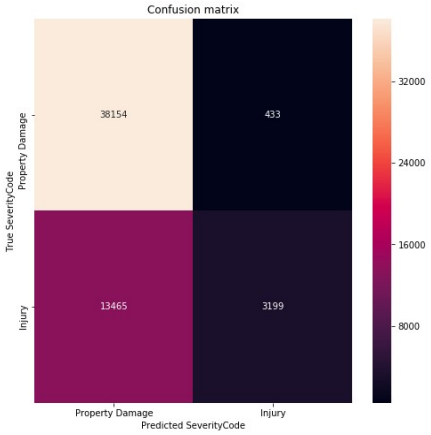
SVM – rbf imbalanced data



	precision	recall	f1-score	support
1	0.75	0.98	0.85	38587
2	0.82	0.23	0.36	16664
accuracy			0.75	55251
macro avg	0.78	0.60	0.60	55251
weighted avg	0.77	0.75	0.70	55251

Log Loss: 24.122253331459294

SVM – linear – imbalanced data



	precision	recall	f1-score	support
1	0.74	0.99	0.85	38587
2	0.88	0.19	0.32	16664
accuracy			0.75	55251
macro avg	0.81	0.59	0.58	55251
weighted avg	0.78	0.75	0.69	55251

Log Loss: 24.122253331459294

Conclusion

Logistic regression with imbalanced data is the best algorithm in this cases. We can predict half accidents in Seattle, saving many lifes when autohorities use this information.