Car Crash Severity Capstone Project

Desiana Nurchalifah

Detecting Car Crash Severity is Important for AI Companies

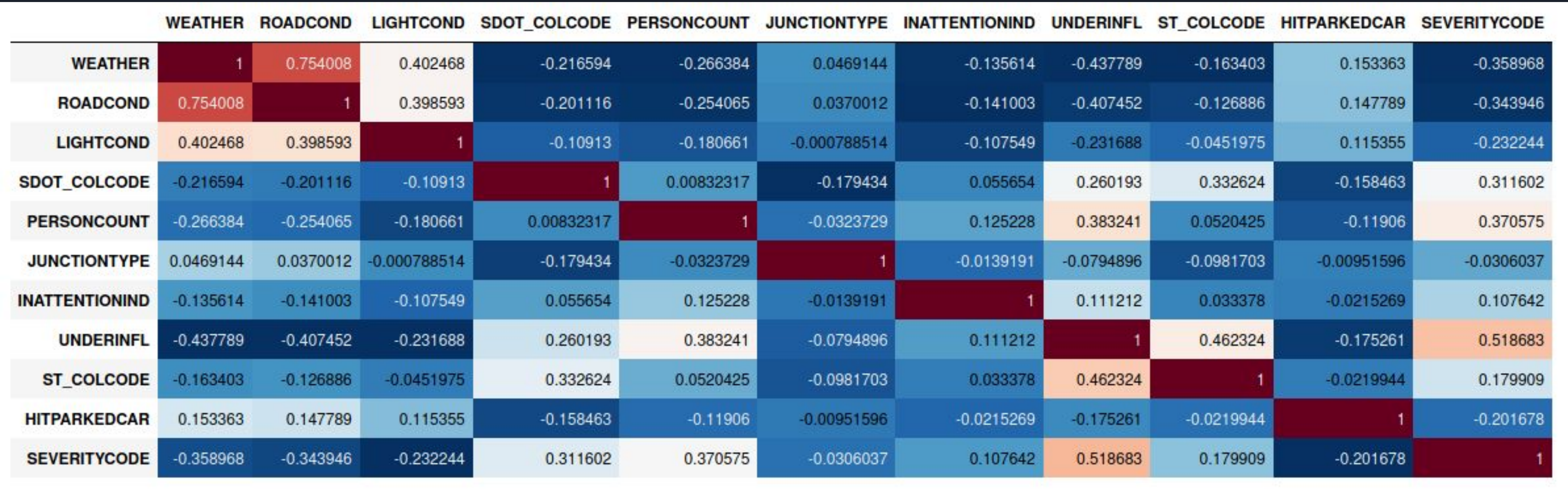
- Growth interest in AI systems, especially Autonomous Driving
- Prediction by governments to automate driving systems (BCG, 2015)
- Safety is still intricate issue, therefore
  - Observe variables that took part in car crash severity from real-life data
  - Comparison of best machine learning methods to classify severity
- Model is also attractive to build an alert system for city infrastructure workers (paramedics, police, firefighter, etc)

Data

- Seattle GeoData. The data is an open data from the Seattle government.[1]
- Data is collected from the year 2004 until present time. (2020)
- The data includes all types of collisions.
- The data consists of 40 independent variables as potential features and 221,144 collection of accidents.
- The dependent variable, **SEVERITYCODE**, contains numbers that correspond to different levels of severity caused by an accident from 0 to 3.
- Among 40 attributes, 10 are chosen to be features.

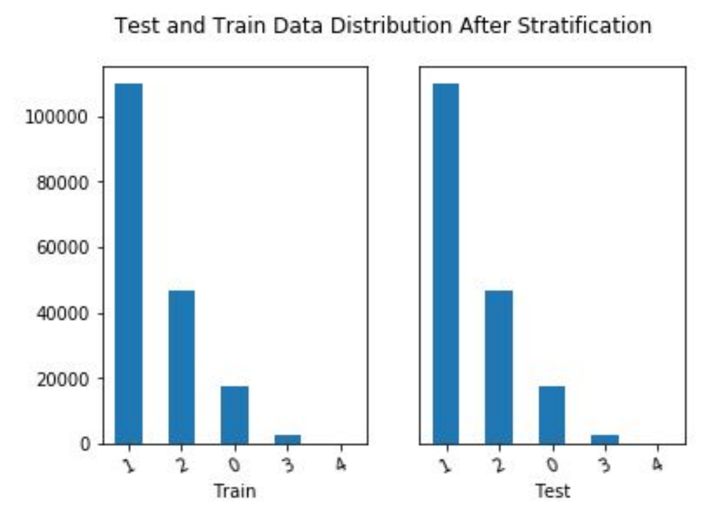
[1]https://data-seattlecitygis.opendata.arcgis.com/datasets/5b5c745e0f1f48e7a53acec63a0022ab_0?geometry=-124.788%2C47.371%2C-119.732%2C48.018

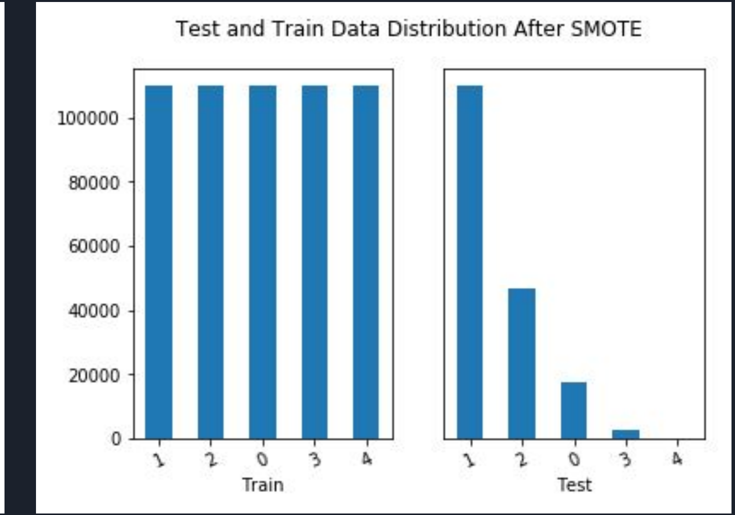
Correlation of Variables



	WEATHER	ROADCOND	LIGHTCOND	SDOT_COLCODE	PERSONCOUNT	JUNCTIONTYPE	INATTENTIONIND	UNDERINFL	ST_COLCODE	HITPARKEDCAR	SEVERITYCODE
WEATHER	1	0.754008	0.402468	-0.216594	-0.266384	0.0469144	-0.135614	-0.437789	-0.163403	0.153363	-0.358968
ROADCOND	0.754008	1	0.398593	-0.201116	-0.254065	0.0370012	-0.141003	-0.407452	-0.126886	0.147789	-0.343946
LIGHTCOND	0.402468	0.398593	1	-0.10913	-0.180661	-0.000788514	-0.107549	-0.231688	-0.0451975	0.115355	-0.232244
SDOT_COLCODE	-0.216594	-0.201116	-0.10913	1	0.00832317	-0.179434	0.055654	0.260193	0.332624	-0.158463	0.311602
PERSONCOUNT	-0.266384	-0.254065	-0.180661	0.00832317	1	-0.0323729	0.125228	0.383241	0.0520425	-0.11906	0.370575
JUNCTIONTYPE	0.0469144	0.0370012	-0.000788514	-0.179434	-0.0323729	1	-0.0139191	-0.0794896	-0.0981703	-0.00951596	-0.0306037
INATTENTIONIND	-0.135614	-0.141003	-0.107549	0.055654	0.125228	-0.0139191	1	0.111212	0.033378	-0.0215269	0.107642
UNDERINFL	-0.437789	-0.407452	-0.231688	0.260193	0.383241	-0.0794896	0.111212	1	0.462324	-0.175261	0.518683
ST_COLCODE	-0.163403	-0.126886	-0.0451975	0.332624	0.0520425	-0.0981703	0.033378	0.462324	1	-0.0219944	0.179909
HITPARKEDCAR	0.153363	0.147789	0.115355	-0.158463	-0.11906	-0.00951596	-0.0215269	-0.175261	-0.0219944	1	-0.201678
SEVERITYCODE	-0.358968	-0.343946	-0.232244	0.311602	0.370575	-0.0306037	0.107642	0.518683	0.179909	-0.201678	1

Handling Imbalanced Data

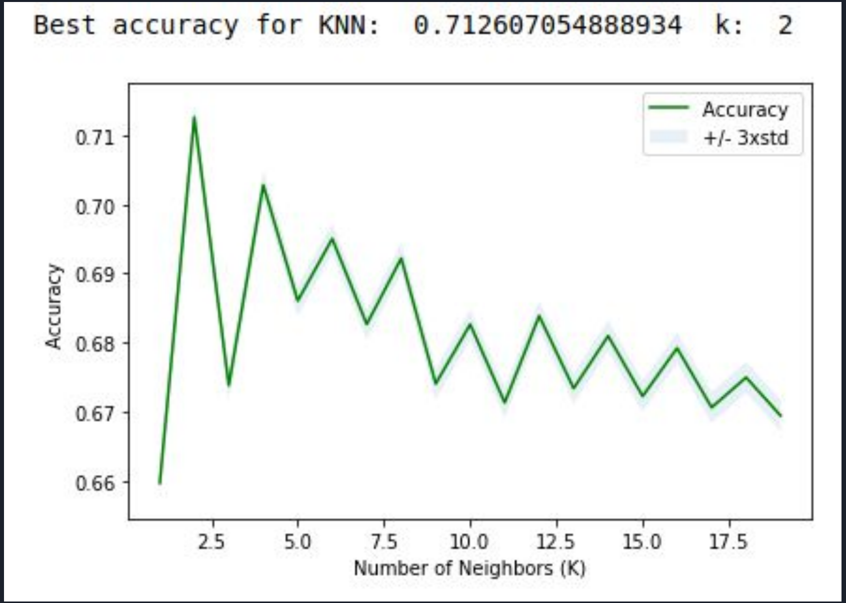




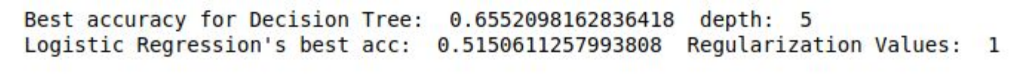
As there are 4 classes of severity type, the sum of each classes are imbalanced, hence to balance the data distributions for both training and test data, stratification is used.
Meanwhile, fairness to classify multi-class data is also imperative. Hence, SMOTE is applied to the training data.

Building Models

All models are previously searched in the space of [1,20] for k, [1,15] for depth and [0.001,0.01,0.1,1,10,100] for regression in logistic regression.



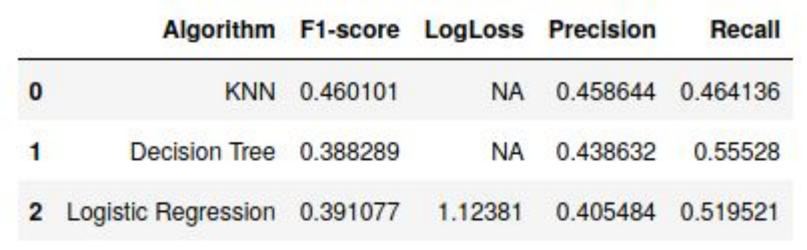
```
Best accuracy for KNN:  0.712607054888934  k:  2
```



```
Best accuracy for Decision Tree:  0.6552098162836418  depth:  5
Logistic Regression's best acc:  0.5150611257993808  Regularization Values:  1
```

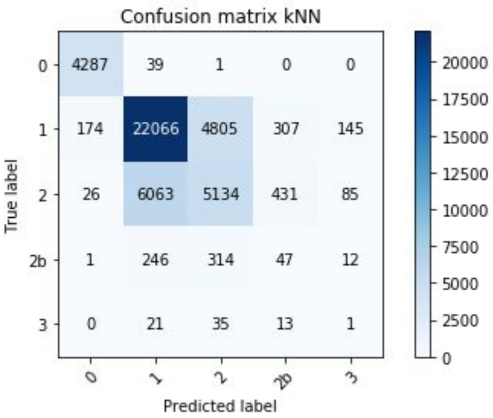
Results

In the results, it shows that among three machine learning methods, KNN excels other methods with only a small difference in recall. Although KNN provides the best performance evaluation, parameter tuning in KNN is computationally exhaustive.



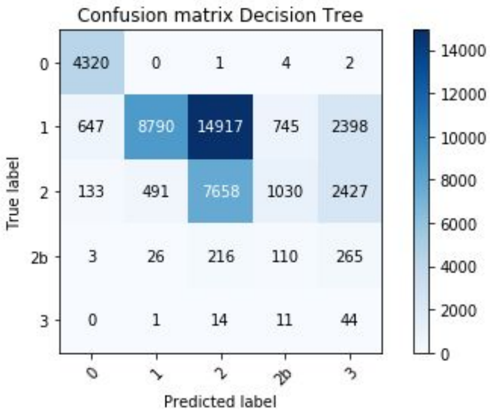
	Algorithm	F1-score	LogLoss	Precision	Recall
0	KNN	0.460101	NA	0.458644	0.464136
1	Decision Tree	0.388289	NA	0.438632	0.55528
2	Logistic Regression	0.391077	1.12381	0.405484	0.519521

Results



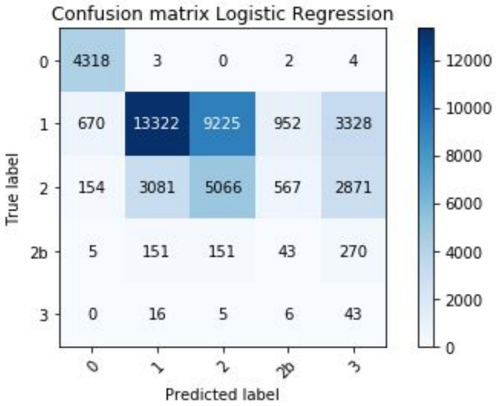
Confusion matrix kNN

	0	1	2	2b	3
0	4287	39	1	0	0
1	174	22066	4805	307	145
2	26	6063	5134	431	85
2b	1	246	314	47	12
3	0	21	35	13	1



Confusion matrix Decision Tree

	0	1	2	2b	3
0	4320	0	1	4	2
1	647	8790	14917	745	2398
2	133	491	7658	1030	2427
2b	3	26	216	110	265
3	0	1	14	11	44



Confusion matrix Logistic Regression

	0	1	2	2b	3
0	4318	3	0	2	4
1	670	13322	9225	952	3328
2	154	3081	5066	567	2871
2b	5	151	151	43	270
3	0	16	5	6	43

- KNN is the safest approach.
- KNN has lowest misclassification value for severe accident (type 3) and low probability accidents (type 1 and type 2).
- The highest total of true label which classified correctly is also achieved by kNN, followed by Logistic Regression and Decision Tree respectively.

Conclusion

- Among 40 attributes which have a potential value in predicting car crash severity, human intoxication factor is still the highest feature importance that takes part in severity of car accidents.
- Among environment conditions, weather is imperative to contribute in car collisions, while location of the collision, such as junctions, are exceptionally insignificant.
- To predict whether collision will affect traffic, with the state of property damage to severity that leads to fatal injuries of the accident, it could be classified effectively using kNN, which had performed better than Decision Tree or Logistic Regression