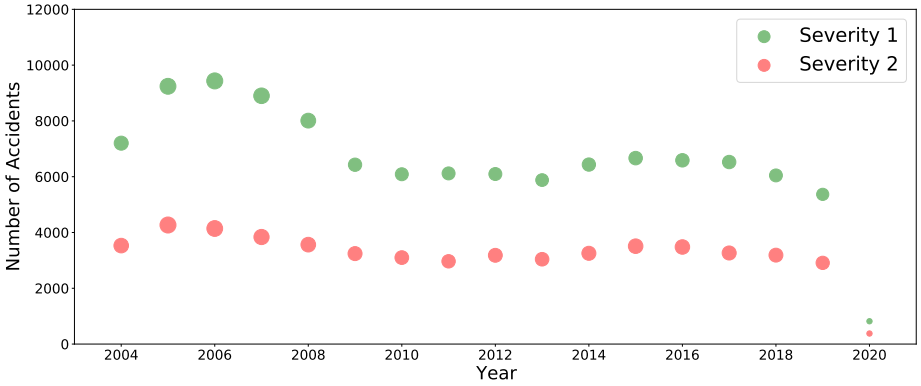
Predicting the Severity of Possible Traffic Accidents in Seattle

Rong Chen

13 October, 2020, Capstone project for IBM Data Science Professional Certificate.

Need to predict the severity of possible traffic accident in real time

- Data shows more than 10000 traffic accidents occur every year in Seattle city!



- With state of the art machine learning algorithms, now it is possible to predict the severity before it occurred even in real time Life can be saved! Be prevenger, not avenger!

Imbalanced: # severity 1 accidents are 2 × # severity 2 accidents

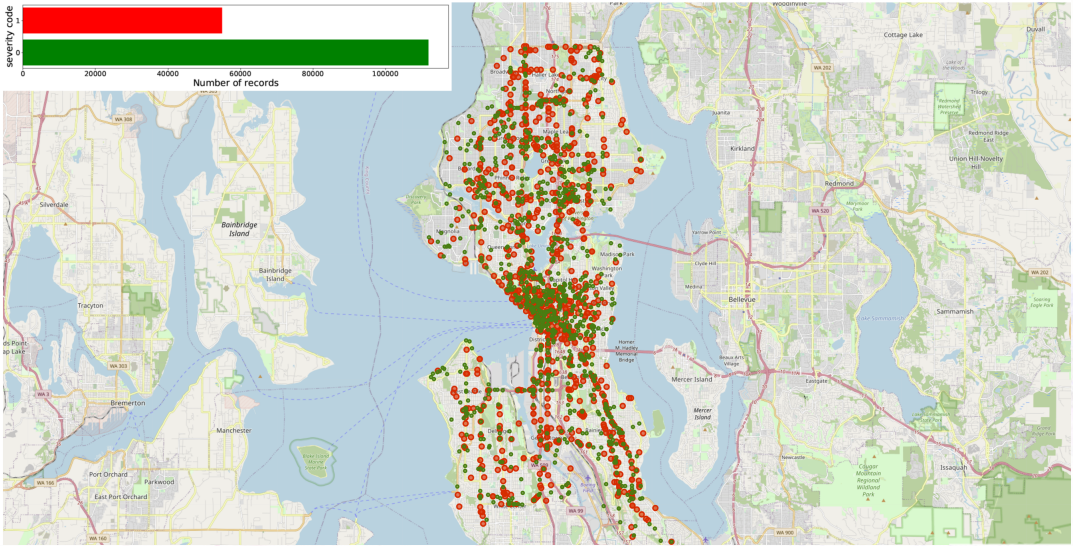
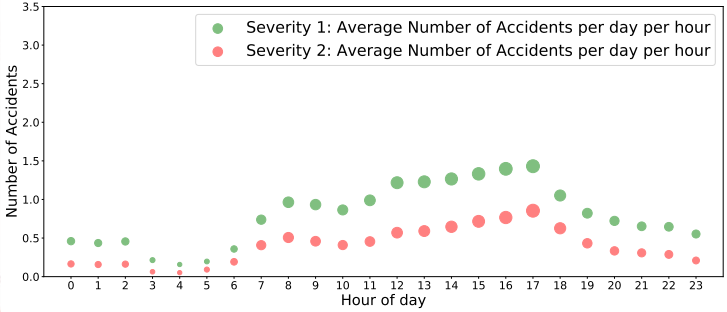


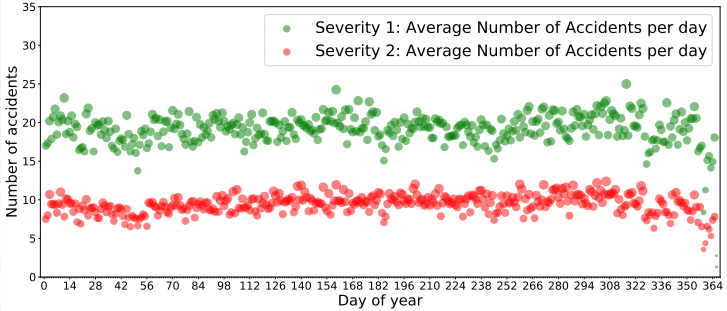
Table: Features in the raw data

Label	Features (37)
SEVERITYCODE	X, Y, OBJECTID, INCKEY, COLDETKEY, REPORTNO, STATUS, ADDRTYPE, INTKEY, LOCATION, EXCEPTRSNCODE, EXCEPTRSNDESC, SEVERITYCODE.1, SEVERITYDESC, COLLISIONTYPE, PERSONCOUNT, PEDCOUNT, PEDCYLCOUNT, VEHCOUNT, INCDATE, INCDTTM, JUNCTIONTYPE, SDOT_COLCODE, SDOT_COLDESC, INATTENTIONIND, UNDERINFL, WEATHER, ROADCOND, LIGHTCOND, PEDROWNOTGRNT, SDOTCOLNUM, SPEEDING, ST_COLCODE, ST_COLDESC, SEGLANEKEY, CROSSWALKKEY, HITPARKEDCAR

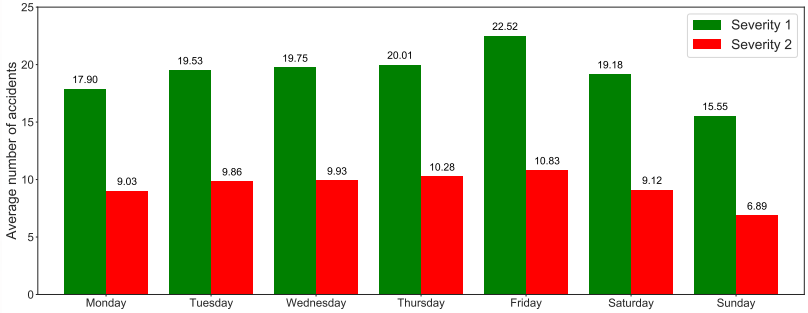
Table: Relevant features for predicting the severity code 1 or 2.

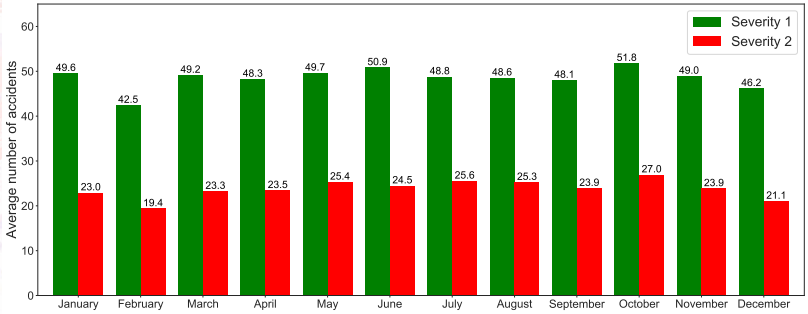
Label	Features (11)
SEVERITYCODE	X, Y, ADDRTYPE, INCDATE, INCDTTM, INATTENTIONIND, UNDERINFL WEATHER, ROADCOND, LIGHTCOND, SPEEDING





- The ratio of the number of severity 1 accidents and severity 2 accidents is around 2:1.
- 10AM to 5PM the number of accidents increases.
- 5PM to 4AM the number of accidents decreases.
- 2AM to 4AM is the 'sleeping' hour of the Seattle city.
- Towards the end of the year, the number of accidents decreases.





- The ratio of the number of severity 1 accidents and severity 2 accidents is around 2:1.
- Monday to Friday the number of accidents increases.
- Friday to Sunday the number of accidents decreases.
- October has the highest number of accidents.
- From October towards the end of the year, the number of accidents decreases.

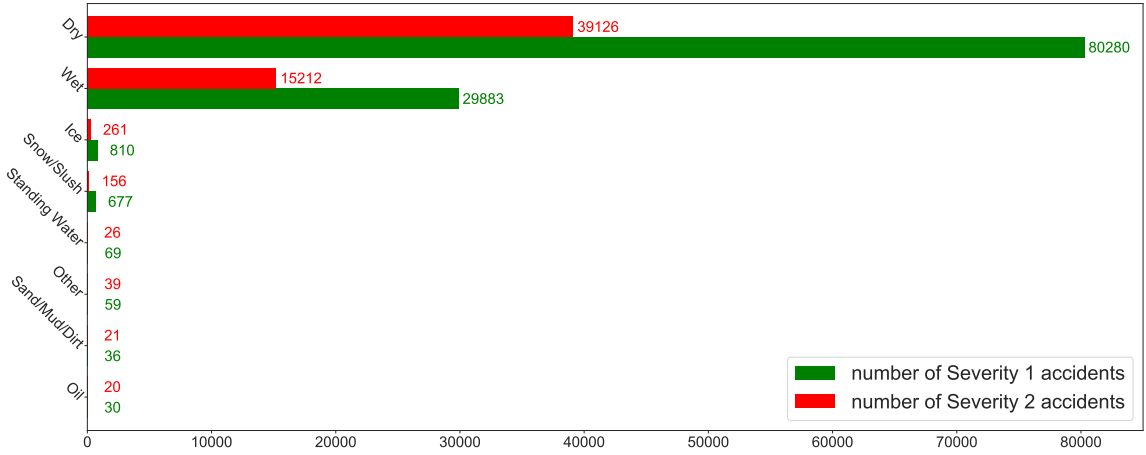


Figure: Number of accident under different road conditions.

- The ratio of the number of severity 1 accidents and severity 2 accidents is around 2:1.
- Under ice, snow/slush conditions, ratio becomes 3:1.

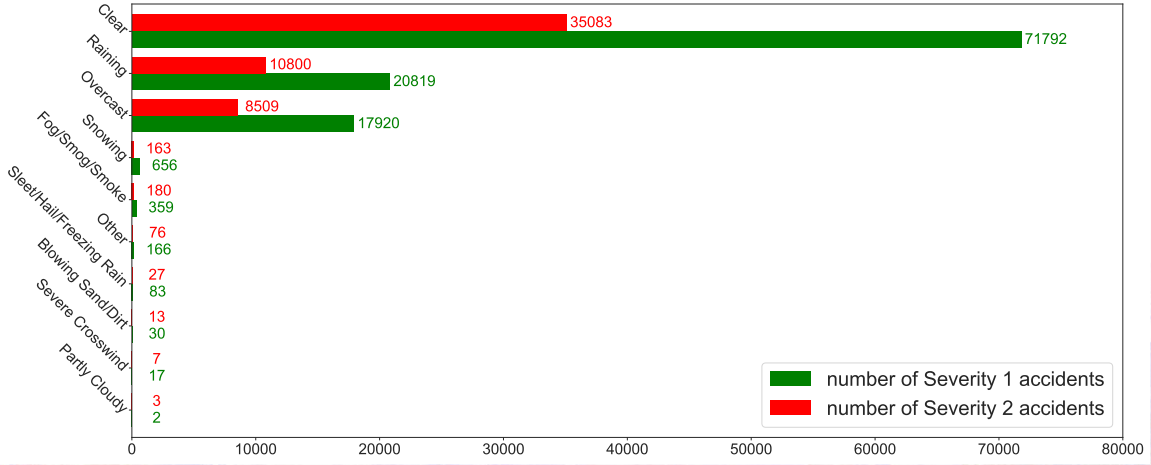


Figure: Number of accident under different weather conditions.

- The ratio of the number of severity 1 accidents and severity 2 accidents is around 2:1.
- Under snowing, sleet/hail/freezing rain conditions, ratio becomes more than 3:1.

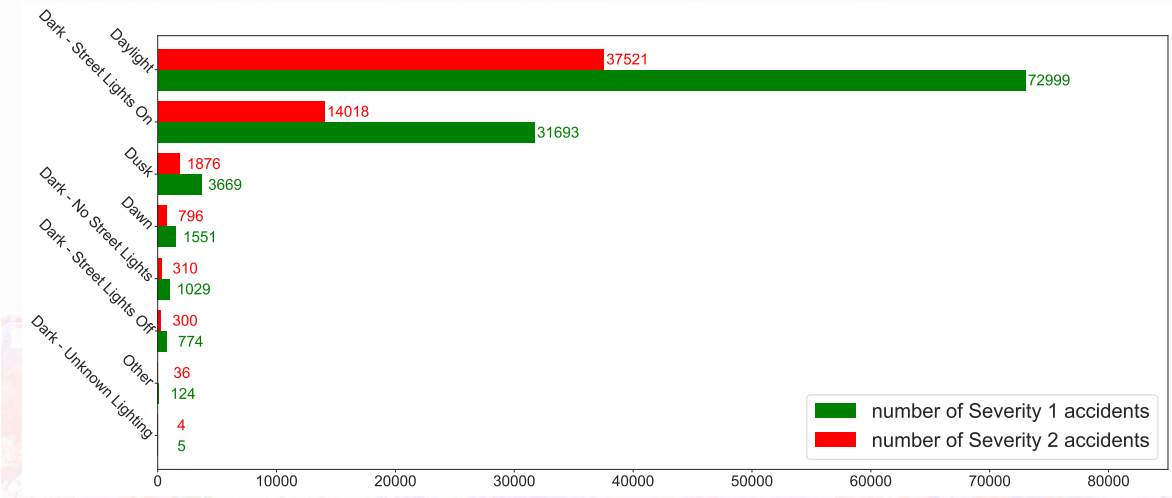


Figure: Number of accident under different light conditions.

- The ratio of the number of severity 1 accidents and severity 2 accidents is around 2:1.
- Under Dark - No steet lights conditions, ratio becomes around 3:1.

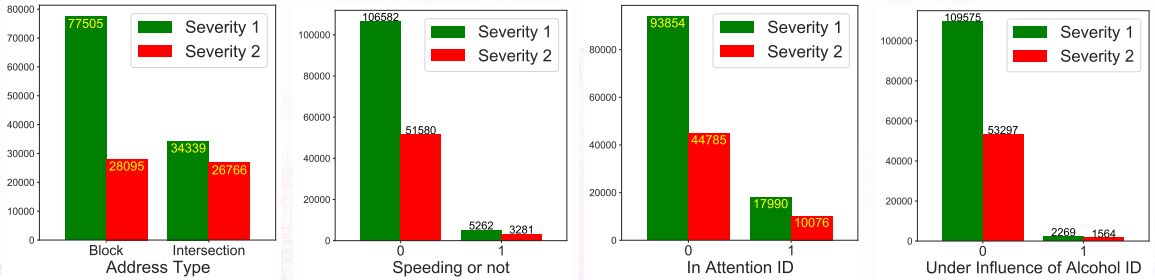


Figure: The number of severity 1 and 2 accidents under different address type, and conditions such as whether speeding or not, paying attention or not, and whether under influence of alcohol or not.

- The ratio of the number of severity 1 accidents and severity 2 accidents is around 2:1.
- When address type is intersection, severity 2 accidents is likely to happen because the ratio becomes nearly 1:1.

Table: Useful features after feature engineering

Label	Features (39)
SEVERITYCODE	INATTENTIONIND, UNDERINFL, SPEEDING, ADDRTYPE_Block, ADDRTYPE_Intersection, WEATHER_Blowing Sand/Dirt WEATHER_Clear, WEATHER_Fog/Smog/Smoke WEATHER_Other, WEATHER_Overcast, WEATHER_Partly Cloudy, WEATHER_Raining, WEATHER_Severe Crosswi WEATHER_Sleet/Hail/Fre WEATHER_Snowing, ROADCOND_Dry, ROADCOND_Ice, ROADCOND_Oil, ROADCOND_Other, ROADCOND_Sand/Mud/Dirt ROADCOND_Snow/Slush, ROADCOND_Standing Wate ROADCOND_Wet, LIGHTCOND_Dark - No St LIGHTCOND_Dark - Stree LIGHTCOND_Dark - Stree LIGHTCOND_Dark - Unkno LIGHTCOND_Dawn, LIGHTCOND_Daylight, LIGHTCOND_Dusk, LIGHTCOND_Other, dayofyear_cos, dayofyear_sin, minuteofday_cos, minuteofday_sin, Longitude_cos, Longitude_sin, Latitude_cos, Latitude_sin

- Take location coordinates and day and time coordinates in to sine and cosine forms.
- One hot engineering is used.
- Feature scaling is used.

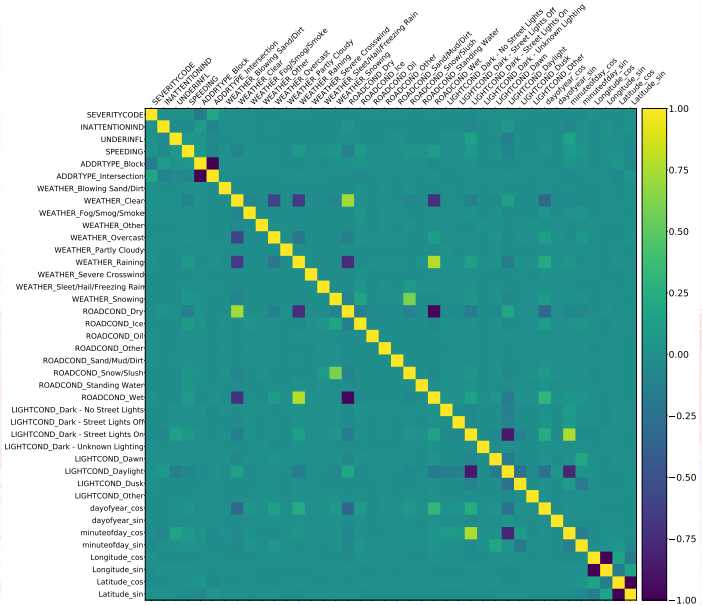


Figure: Correlation matrix for the severity code and the features used.

The address type, block or intersection, has more correlation with severity code than other features.

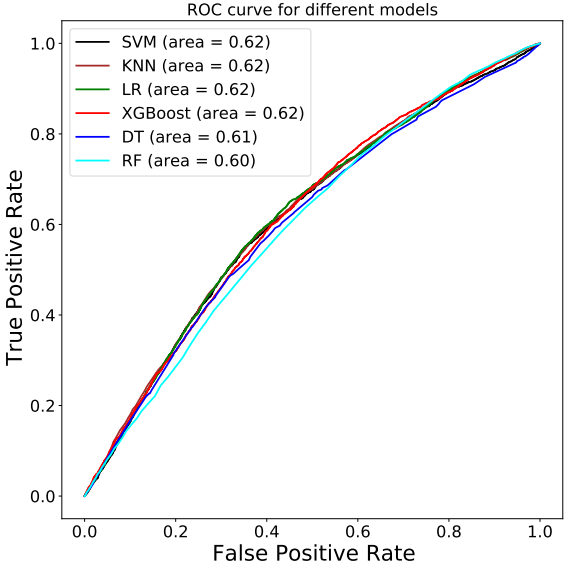
Supervised machine learning models for classification

In order to predict the severity code 1 or 2 for the possible traffic accident in Seattle city, we use the following models to train the historical data:

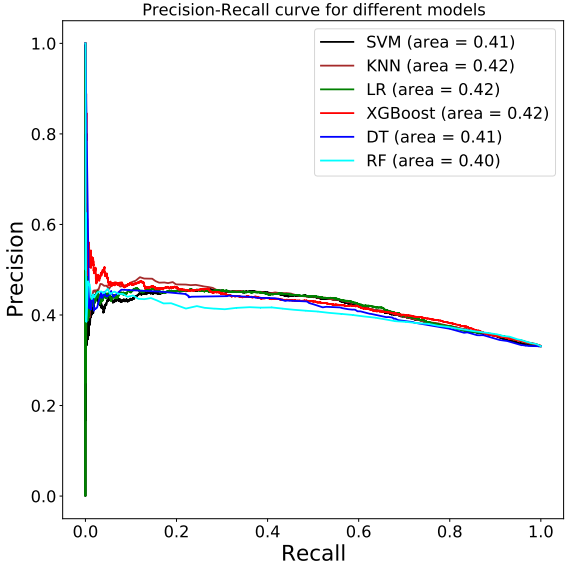
- Logistic regression (LR).
- Decision three (DT).
- Extreme gradient boosting (XGBoost).
- Random forest (RF).
- K-nearest neighbor (KNN).
- Support vector machine (SVM).

We test:

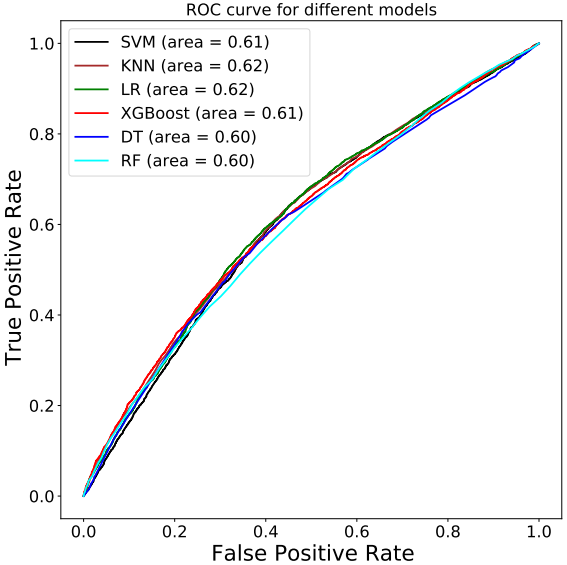
- Jaccard score.
- F1-score.
- Accuracy.
- Receiver operator curve (ROC) and the area under curve (AUC).
- Precison/recall curve and AUC.
- Number of severity 1 and severity 2 accidents predicted.



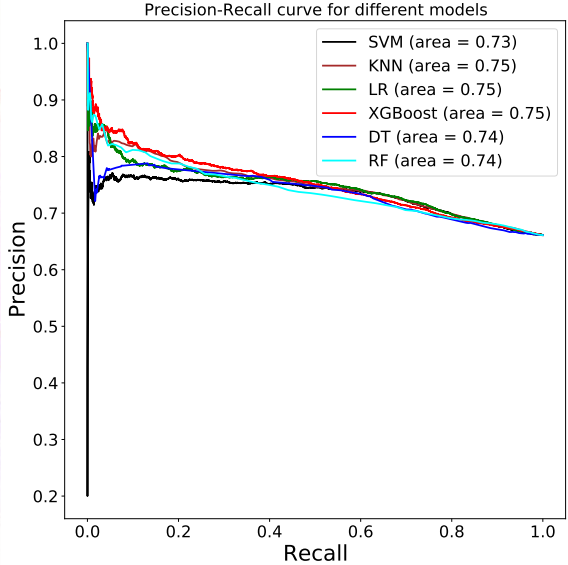
(a) Severity 2 as positive and 1 as negative.



(b) Severity 2 as positive and 1 as negative.



(a) Severity 1 as positive and 2 as negative.



(b) Severity 1 as positive and 2 as negative.

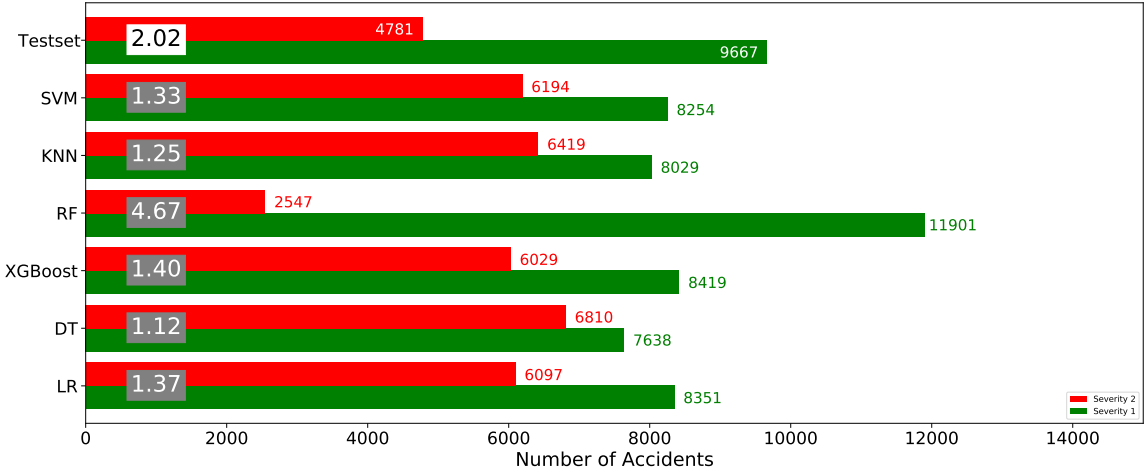


Figure: Model predicted number of severity 1 and 2 accidents and the theoretical number of accidents in the testset.

Table: The accuracy of the built model using different evaluation metrics

Model	Training set			Test set		
	Jaccard index	F1-score	Accuracy	Jaccard index	F1-score	Accuracy
SVM	0.32	0.62	0.61	0.32	0.62	0.61
KNN	0.42	0.60	0.60	0.32	0.62	0.61
LR	0.41	0.60	0.60	0.32	0.62	0.61
XGBoost	0.46	0.63	0.63	0.32	0.62	0.61
DT	0.46	0.63	0.63	0.32	0.61	0.59
RF	1.00	1.00	1.00	0.19	0.62	0.65

- All the models have skills.
- LR, XGBoost, SVM, and KNN performs relatively good, accuracy more than 60%.
- RF suffers over fitting issue.
- DT has relatively low accuracy.

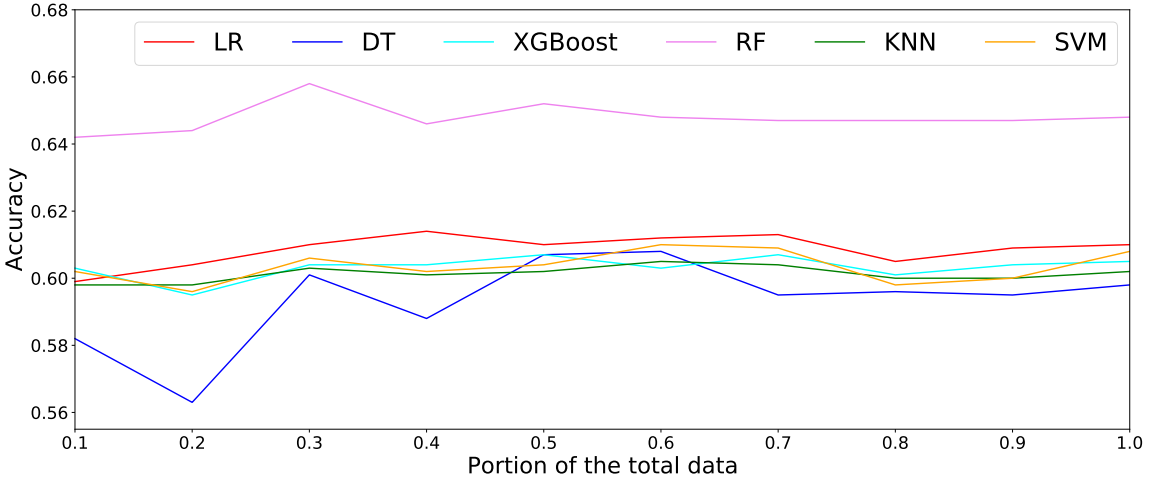


Figure: the testing accuracy .vs. the amount of training data we used

- None of the model shows rising tendency of accuracy with the increasing amount of the data.
- It is not the more data you train the better results you get.

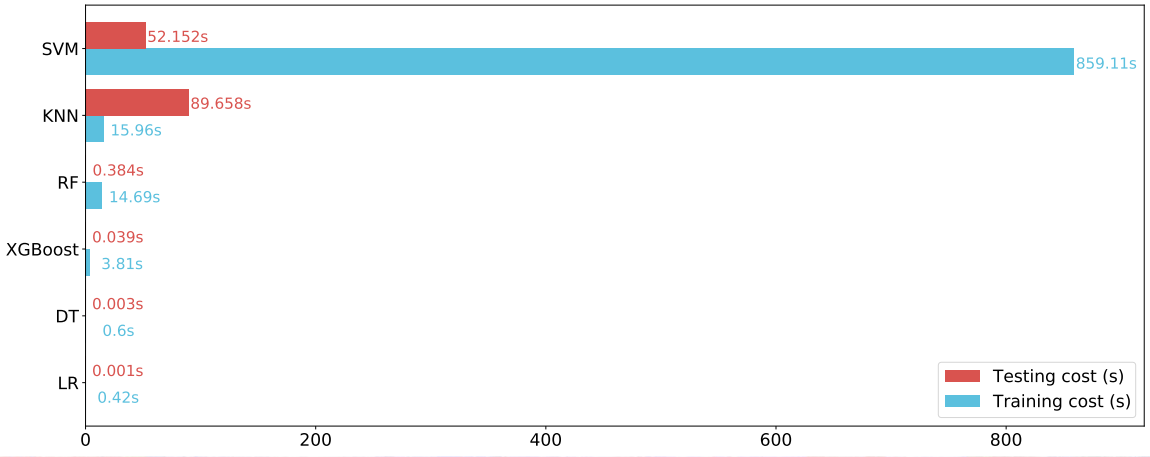


Figure: Training and test time cost for different models based on about 80000 examples (50% of the total data) including 64000 training examples and 16000 testing examples.

- LR, DT and XGBoost are cheap.
- SVM is expensive to train.
- KNN is expensive in making prediction.

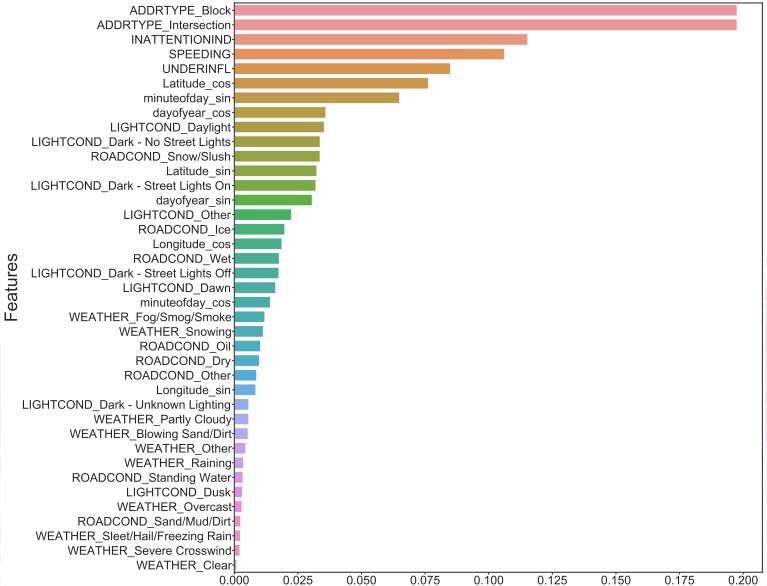


Figure: Importance of the features (LR).

Top features for predicting severity 1 and severity 2 accidents:

- Address type: block or intersection
- Whether the driver is in attention or not.
- Whether the driver is in speeding or not.
- Whether the driver is in under influence or not.
- Location.
- Time of the day.
- Day of year.
- light condition.
- Road condition.
- Weather condition.

Summary and Outlook

Summary

- Trained LR, SVM, XGBoost, KNN, DT, and RF models and predicted the possible severity 1 and 2 accidents in Seattle city with reasonable accuracy.
  - LR, SVM, XGBoost, and KNN are relatively good.
  - DT has lower accuracy and RF suffers overfitting.
- Recommandations for software developer who makes prediction/slarming software
  - For powerful computing devices, LR, SVM, XGBoost, and KNN can provide real time predictions.
  - For cell phone Apps, LR is the best choice for real time predictions.
- Recommandations for drivers, passengers, and pedestrians
  - Be very careful at intersection area where severity 2 accidents is more likely to happen.
  - From noon to 5PM during work days accidents are more likely to happen so be aware.
  - Be focused during driving, do not drink alcohol, do not speeding.
  - Be aware of the light, road, and weather conditions.

Outlook

- Adding more feature can help, such as knowing which direction and what speed the car may be hit.
- Develop more suitable algorithms, neural network may help.