CAR ACCIDENT SEVERITY

IBM Specialization – Capstone Project

Introduction/Business Problem

The goal of this project is to predict the severity of a traffic accident. Car accidents are undesirable and dangerous for all traffic participants. Information about road conditions and possibility of one getting into car accident and its intensity could be signal for changing a traffic route. This, along with driving safely at all time, could save many lives caused by traffic disasters. Therefore, prediction of traffic accident severity is one of the main things that improve traffic management process. Since car accidents are unexpected, recognition of most important influences might be a substantial key for improving traffic safety. Said that, it is obvious that WHS consultant, facilitators, traffic consultants, safety officers and other people involved in traffic management process might find this kind of project quite useful.

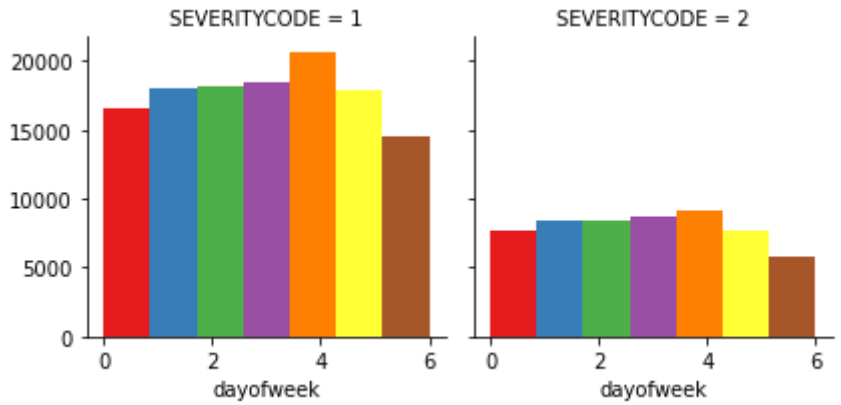
Data

Data used in this project are data about trafic collisions recorded in Seattle area. All data are provided by Seattle Police Department and recorded by SDOT Traffic Management Division, Traffic Records Group since 2004. Data could be downloaded at https://s3.us.cloud-object-storage.appdomain.cloud/cf-courses-data/CognitiveClass/DP0701EN/version-2/Data-Collisions.csv

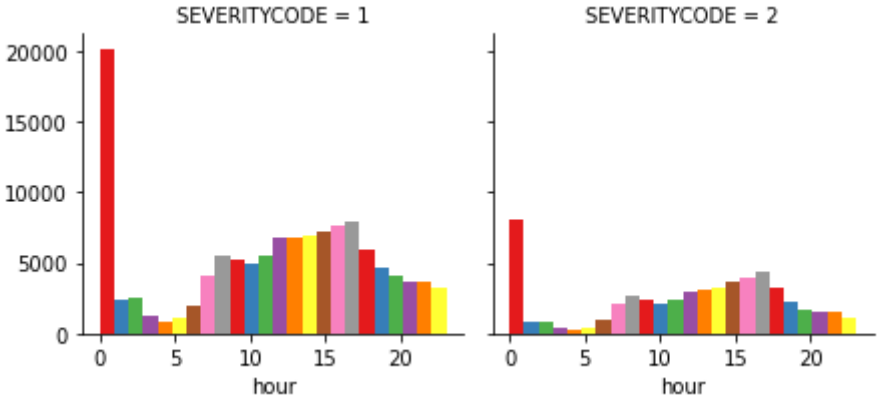
In total, there are 37 attributes. Target of the data is severity which describes the fatality of an accident (variable SEVERITYCODE). Label Y, 'SEVERITYCODE' variable has two outputs in our data set: 1 - prop damage and 2 – injury.

While reviewing data in order to choose potential features set, fields of long unique descriptions, fields representing unique ID data related to the recordings of the police or some other institution and data that are recorded after collision happened are disregarded. Data that might contribute to determining accident severity might include collision address type (ADDRTYPE), date and time (INCDTTM), category of junction (JUNCTIONTYPE), whether or not a driver involved is under the influence of drugs or alcohol (UNDERINFL), weather conditions (WEATHER), condition of the road (ROADCOND) and light conditions (LIGHTCOND).

Address type include two values – Block (0) and Intersection (1). Variable about date and time of recorded collision (INCDTTM) are transformed into two additional variables ‘HOUR’ and ‘DAYOF WEEK’. Someone might find this strange, but there are there are no more accidents during the weekend, numbers but almost uniformly spread through the whole week.



Additional information could be found with 'hour' variable since there are evidently more accidents during some hours (around midnight and rush hours) than the others (around 5-6 am), even though there are no distinctiveness between severity of an accident. So, added variable 'HOUR' is transformed into 3 groups as following: 1 – time (hours) of high risk of collision, 2 – time of moderate risk of collision, 3 – time of low risk of collision.



Variables 'ADDRTYPE' and 'JUNCTIONTYPE' both represent category of junction at which collision took place, so there is no reason to use both variables in features set. Since 'JUNCTIONTYPE' represents more complex division, this variable could add more information to the model.

After cleaning dataset, converting categorical features to numerical values, creating dummy variables when needed, as well as the normalizing the data, we obtain following dataset:

FEATURE	VALUE	DESCRIPTION
hour	1	Time/Hour of Low Risk of Collision
	2	Time/Hour of Moderate Risk of Collision
	3	Time/Hour of High Risk of Collision
JUNCTIONTYPE	1	Ramp junction or unknown
	2	At Intersection (but not related to intersection)
	3	Driveway Junction
	4	Mid-Block (but intersection related)
	5	At Intersection (intersection related)
	6	Mid-Block (not related to intersection)
UNDERINFL	0	Involved Driver Not Under Influence of Drugs or Alcohol.
	1	Involved Driver Under Influence of Drugs or Alcohol.
WEATHER	0	Clear
	1	Unknown or Other
	2	Partly Cloudy
	3	Overcast
	4	Severe Crosswind
	5	Raining
	6	Snowing
	7	Blowing Sand/Dirt
	8	Fog/Smog/Smoke
	9	Sleet/Hail/Freezing Rain
ROADCOND	0	Dry
	1	Unknown or Other
	2	Sand/Mud/Dirt
	3	Standing Water
	4	Snow/Slush
	5	Wet
	6	Oil
	7	Ice
LIGHTCOND	0	Daylight
	1	Unknown or Other
	2	Dawn
	3	Dusk
	4	Dark - Street Lights On
	5	Dark - Unknown Lighting
	6	Dark - Street Lights Off or No Street Lights

Methodology

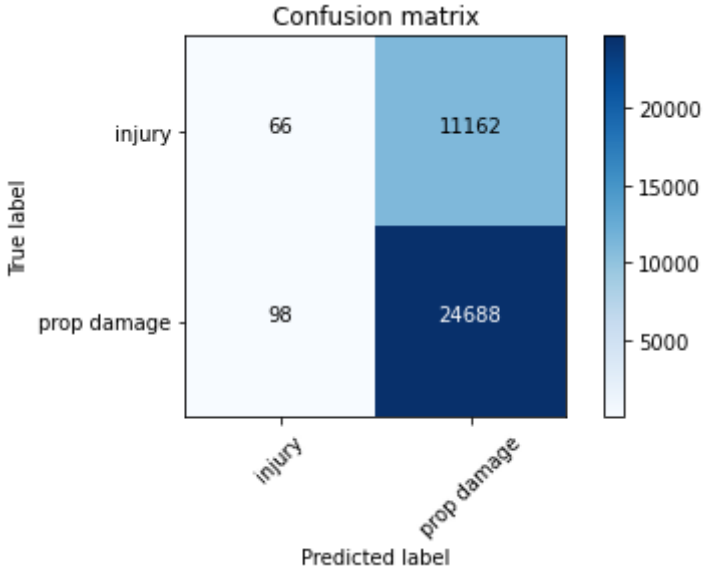
Label Y, 'SEVERITYCODE' variable has two outputs in our data set: 1 - prop damage and 2 – injury. A Linear Regression in this case would not be useful in producing appropriate classification since the label has binary outcome. That is why a Logistic Regression is more appropriate. Logistic Regression relaxes some of the assumptions made by a linear regression model. The relationship between the dependent variable and the predictors does not have to be linear, the predictors do not have to be normally distributed and also heteroscedastic variances are not needed. Logistic Regression produces probabilities between zero and is used to predict the probability of an accident's severity.

Another good algorithm for classification problems tahat is worth considering is KNN (K-nearest neighbours). It considers the 'K' Nearest Neighbors (points) when it predicts the classification of the test point. We should keep in mind that it is very important to consider the value of K.

For comparison, SVM (Support Vector Machine) algorithm is also considerd. SVM take cares of outliers better than KNN. In this case, when training data is much larger than no. of features(m>>n), KNN is better than SVM. We should test that.

Results

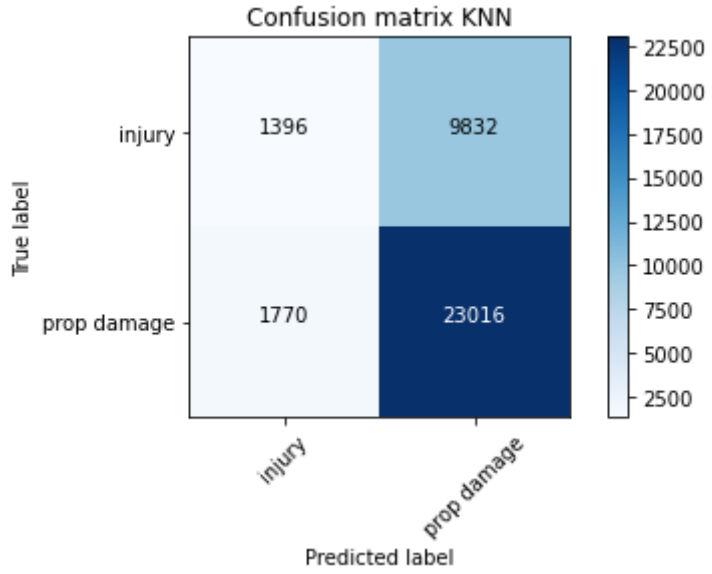
Results for Logistic Regression are shown below. As one could see, model is quite good predictor of *prop damage (1)*, but not so well when predicitng *injury(2).* Logarithmic Loss which quantifies the accuracy of a classifier by penalising false classifications is 0.61. That implies that uncertanty of the model is quite high.



	precision	recall	f1-score	support
1	0.69	1.00	0.81	24786
2	0.40	0.01	0.01	11228
micro avg	0.69	0.69	0.69	36014
macro avg	0.55	0.50	0.41	36014
weighted avg	0.60	0.69	0.56	36014

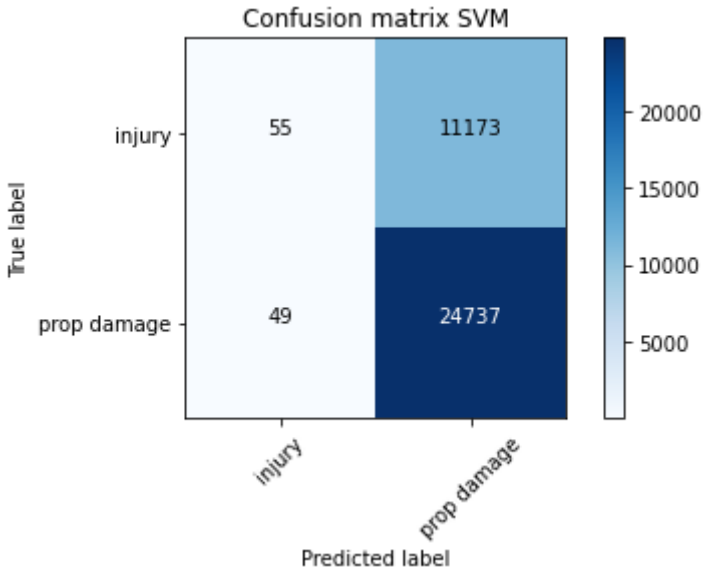
When it comes to KNN algorithm, results are somewhat better. K was selected by elbow method. Elbow method is one of the most popular methods to determine this optimal value of K.

Confusion matrix and accuracy metrics are shown below. As mentioned, results are a bit better than those given by SimpleLogistic Regresion.



	precision	recall	f1-score	support
1	0.70	0.93	0.80	24786
2	0.44	0.12	0.19	11228
micro avg	0.68	0.68	0.68	36014
macro avg	0.57	0.53	0.50	36014
weighted avg	0.62	0.68	0.61	36014

SVM works by mapping data to a high-dimensional feature space so that data points can be categorized, even when the data are not otherwise linearly separable. As one could predict, results are not quite impressive here. Reason could lie in fact that our chosen X set (features set) is not high dimensional.



	precision	recall	f1-score	support
1	0.69	1.00	0.82	24786
2	0.53	0.00	0.01	11228
micro avg	0.69	0.69	0.69	36014
macro avg	0.61	0.50	0.41	36014
weighted avg	0.64	0.69	0.56	36014

Comparison between methods are reported below. Jacard similarity coefficient (index) and F1 score are used as accuracy measures.

	Jaccard	F1-score
KNN	0.68	0.61
SVM	0.69	0.56
LogisticRegression	0.69	0.56

Discussion

Although it is reasonable to expect that accidents occur more often at some parts/hours of the day than the others, it is very noticeable that over 15% of accidents occur at 00-01 am. An hour earlier or later that percentage drops to 2%. Whether it was a coincidence or a mistake, it would be worth examining the reason. Contacting the provider and exploring ther explanation would result in some new findings. Those findings should be incuded in the model, possibly improving it.

When an accident occurs, a police investigation includes assesment and recording of the speed of involved vehicles. Similar publicly available studies include a variable representing the speed above the allowed speed limit. I think that size should definitely be provided and the availability of such data would improve the model.

Conclusion

The main aim of this project was to apply several methods to classify the severity of an injury in car accidents in Seattle, WA area and compare which model makes the most accurate predictions about traffic injury severity. Severity of an accident is labeled by two values - proper damage and injury. Since the binary output, a Simple Logistic Regression in this case was the most appropriate idea to start. Given the data structure, K-nearest Neighbours and Support Vector Machine algorithms were also taken into account. All models had some troubles classifying injury-class. KNN was somewhat more succesful than the other models.

The model has potential and I think it would give more concrete results and be more of a use if data and issues addressed in „Discussion“ part could be acquired and resolved.