Car Accident Severity Prediction Model

Introduction

Car accidents could have severe consequences on our daily lives. By understanding the main factors of the severity of car accidents, we drivers could better plan and manage driving habit to avoid or reduce unnecessary risks. This will also benefit the insurance and the entire society.

The goal of the analysis is intended to identify relative factors and build a quantitative model to predict the severity of the car accidents. The data were collected for the accidents occurred from 2004 to 2020 in Seattle area. The target variable is the severity, which has two levels:

- Property Damage Only Collision
- Injury Collision

As the target variable represents the severity of the accidents with two levels, the classification model would be appropriate for this problem. Applicable models include:

- K-nearest neighbor
- Support Vector Machine
- Decision Tree
- Random Forest
- Logistic Regression

Data

The dataset, provided by Coursera, consists of car accidents data collected from Seattle area from 2004 to 2020. Various attributes of each accidents represent the road condition, weather, location, vehicles and persons involved. Obviously certain attributes relate more closely to the severity of the accidents while others play a lesser role.

Data Quality and Missing Data

The dataset consists of 194673 records with 38 fields. The target variable is SEVERITYCODE with the following distribution:

SEVERITYCODE	SEVERITYDESC	Record Count
1	Property Damage Only Collision	136,485
2	Injury Collision	58,188

According to meta data, certain fields represent the identification of the incidence, which do not have direct relationship to the car accidents, such as 'X', 'Y', 'OBJECTID', 'INCKEY', 'COLDETKEY', 'REPORTNO'

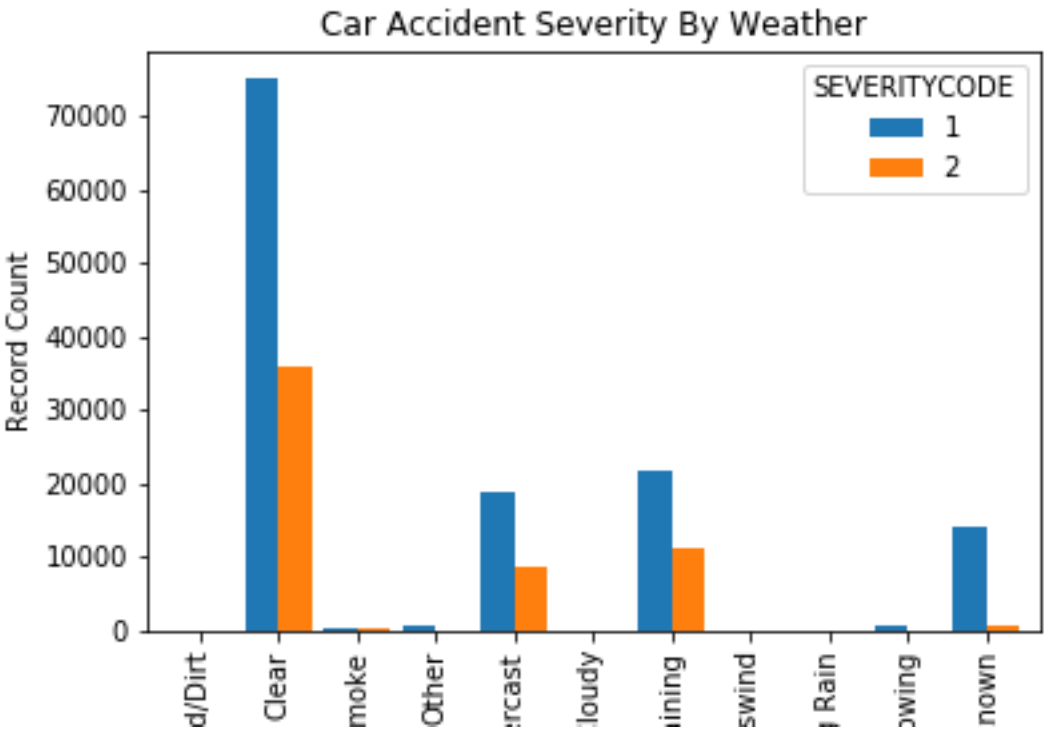
and 'STATUS'. X and Y are longitude and latitude of the location. Additionally, a few fields have significant number of missing values, and they will be excluded from the modeling exercise.

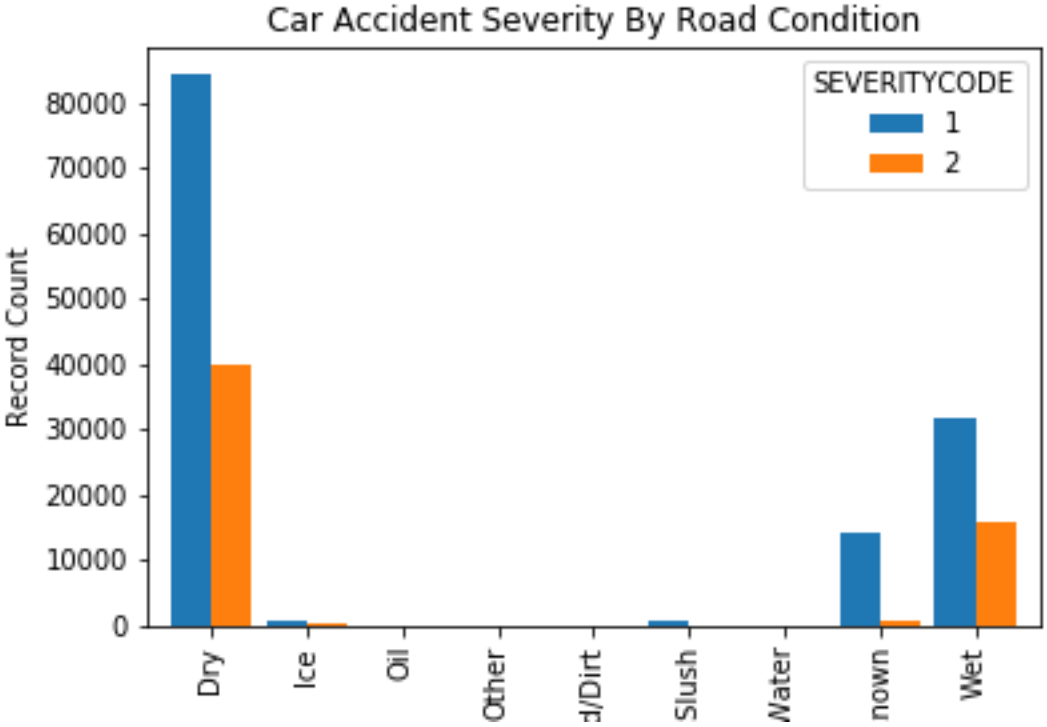
```
Field Name       Percent of Missing Value
-----------------------------------------
INTKEY                 66.57%
EXCEPTRSNCODE          56.43%
EXCEPTRSNDESC          97.10%
INATTENTIONIND         84.69%
PEDROWNOTGRNT          97.60%
SDOTCOLNUM             40.96%
SPEEDING               95.21%
```

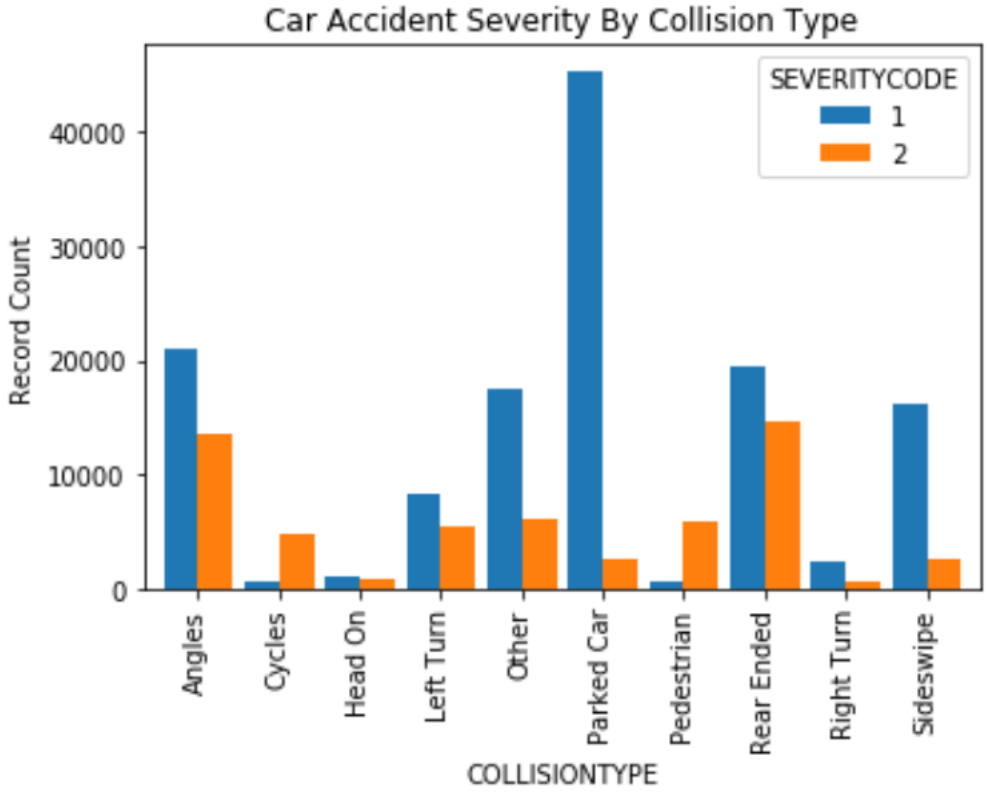
Exploratory analysis and visualization

The data exploration will help identify the variables and uncover relations between independent and target variables.

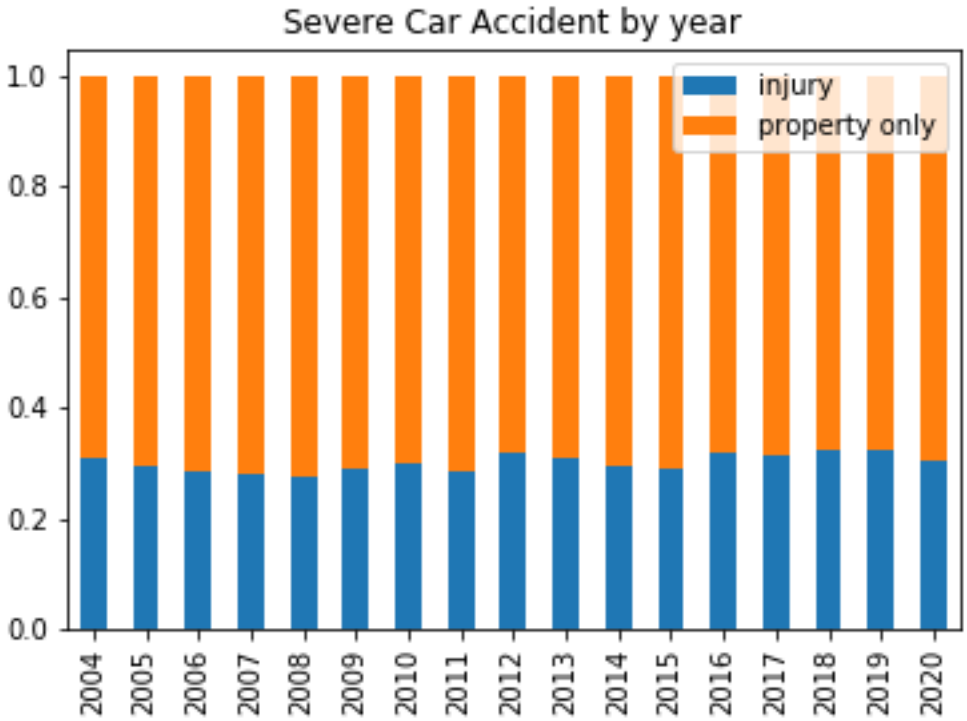
The “WEATHER”, “ROADCOND”, “LIGHTCOND” are categorical variables, and they have some predictive power in relations to the severity of car accidents.



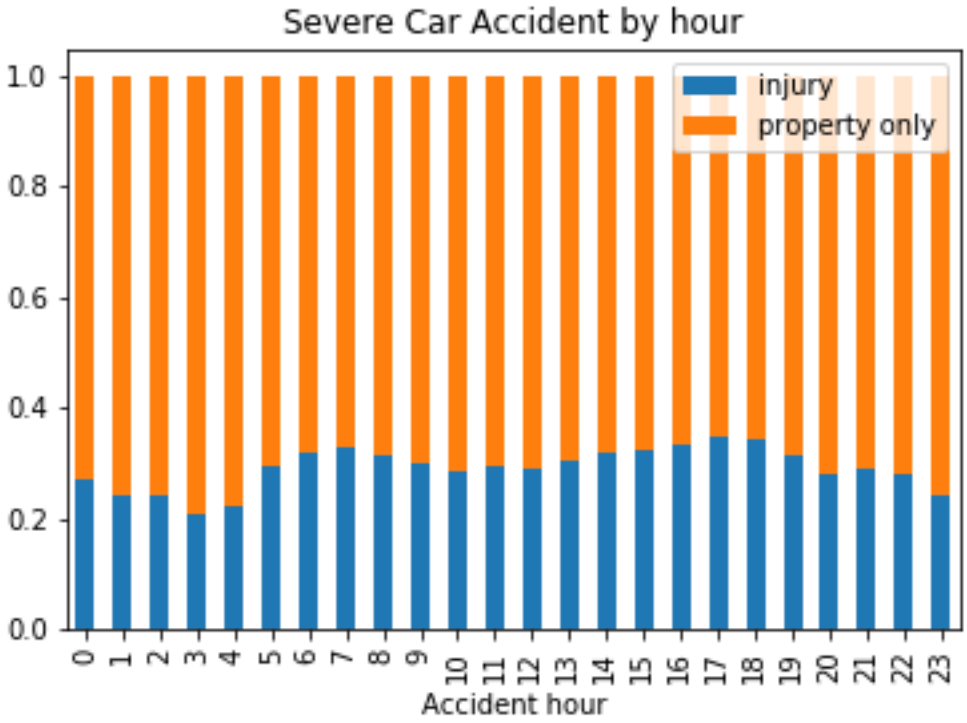

Car Accident Severity By Road Condition
SEVERITYCODE
1
2
Record Count
80000
70000
60000
50000
40000
30000
20000
10000
0
Dry
Ice
Oil
Other
Slush
Wet


Car Accident Severity By Collision Type
SEVERITYCODE
1
2
Record Count
40000
30000
20000
10000
0
Angles
Cycles
Head On
Left Turn
Other
Parked Car
Pedestrian
Rear Ended
Right Turn
Sideswipe
COLLISIONTYPE

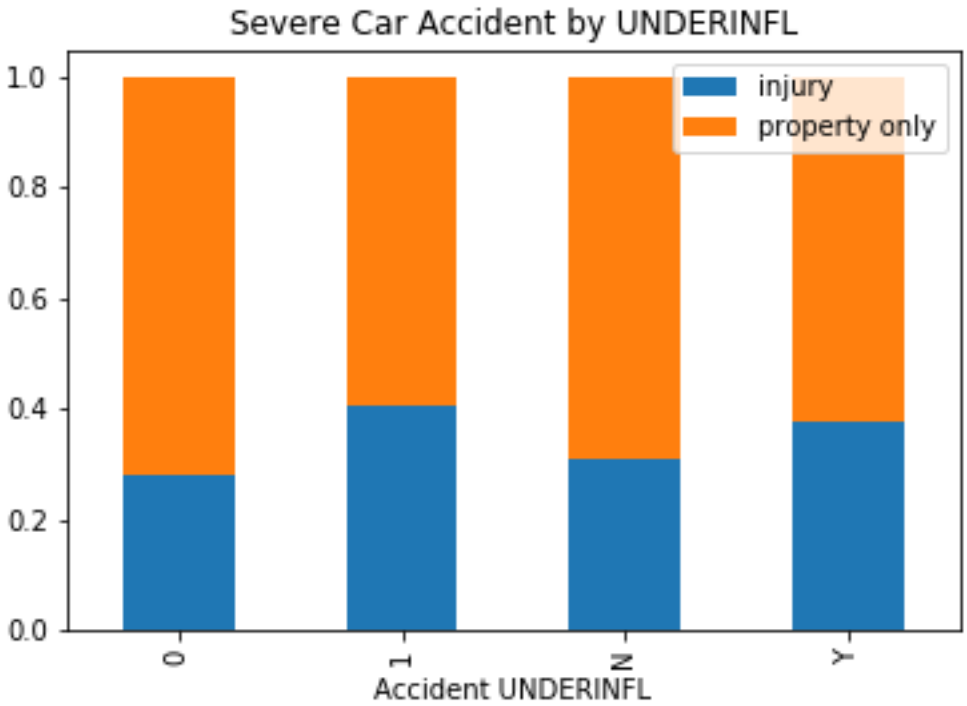
There is no distinct long-term trend in terms of accident severity, as shown below:



However, the time of day does play a role in determining the severity of the accidents, where morning and evening traffic hours tend to observe more injuries.



UNDERINFL has two sets of codes representing if the driver is under influence or not. The distribution is similar, thus combining them.



Data Conversion for Modeling

In preparation to build a predictive model, data cleansing is performed as follows:

1) Categorical variables are converted to on-hot coding for modeling purpose.
2) New features were extracted from 'INCDTTM' field to create new variables representing traffic hours.
3) Selected features include: 'COLLISIONTYPE', 'PERSONCOUNT', 'PEDCOUNT', 'PEDCYLCOUNT', 'VEHCOUNT', 'UNDERINFL', 'WEATHER', 'ROADCOND', 'LIGHTCOND'.
4) Remove rows with missing values. Approximate 2.8% of records were removed.

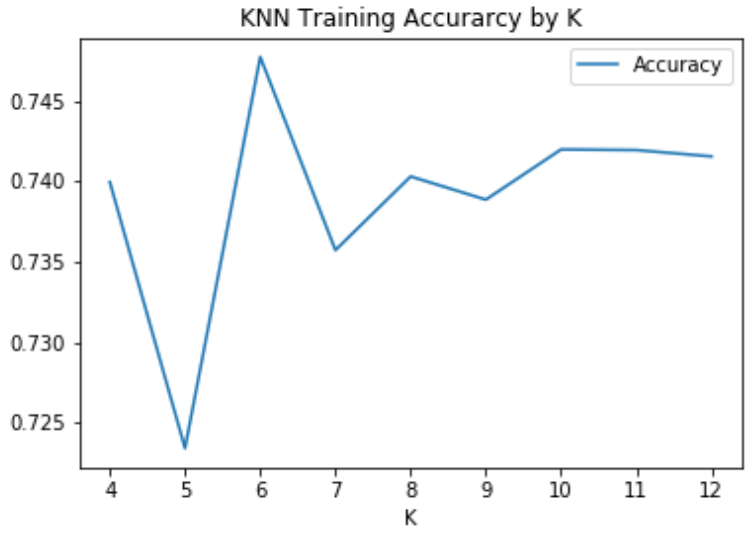
Modeling Methodology

Multiple classification models were fitted for this problem. K-nearest neighbor (KNN) was a non-parametric model that relies on the similarity of the features to predict the outcome. Decision tree and random forest can handle non-linearity well and produce easily interpretable result. Support vector machine (SVM) and logistic regression can provide additional insight on the effect of each variable. Python Scikit learn package provides easy implementation on all these models.

The data set was split into training and testing dataset to facilitate the model selection and evaluation. All models were built by the training dataset and then the performances were separately evaluated on the testing dataset.

K-nearest-neighbor (KNN)

Multiple Ks were tested to select the best K value based on the training data. The accuracy measures for different K are illustrated below. K was set at value of 6 for the final model.




```
KNN model = KNeighborsClassifier(n_neighbors = 6)
```

Decision Tree

The decision tree model was built based on "entropy" criterion.

```
Decision Tree Model = DecisionTreeClassifier(criterion='entropy', max_depth=4)
```

Support Vector Machine (SVM)

The SVM model was built using kernel function of kbf.

Logistics Regression

The logistics classification model was built with regulation parameter set at 0.01.

Random Forest

The random forest model was built using default setting and max depth of 3.

All these models were trained using training data set and multiple performance measures were applied to each model for evaluation.

Results

The models yield the following metrics on the training data set:

Table 1 - Training Data Performance

Model	Precision	Recall	F1-score	Accuracy
KNN	0.768	0.915	0.835	0.748
Decision Tree	0.739	0.990	0.846	0.748
SVM	0.749	0.975	0.847	0.755
LogisticRegression	0.753	0.965	0.846	0.755
Random Forest	0.738	0.990	0.846	0.748

From the above, the performances are similar among the five models. The model selection would depend on the intended purpose of the model output – whether it is more desirable to have higher accuracy or sensitivity.

All the models were evaluated on the unseen test data set and results are as follows:

Table 2 – Test Data Performance

Model	Precision	Recall	F1-score	Accuracy
KNN	0.763	0.905	0.828	0.737
Decision Tree	0.741	0.989	0.847	0.751
SVM	0.752	0.974	0.849	0.758
LogisticRegression	0.756	0.964	0.847	0.757
Random Forest	0.741	0.989	0.847	0.751

The models delivered similar level of performance on the test data as the training data, which indicate that the models did not suffer from the over-fitting problem.

Discussion

All the models only achieved around 75% accuracy, which might not be good enough to make reasonable predictions from the given set of predictors. Other factors might need to be explored in order to improve the model performance. Additionally, some of the predictors may correlate to each other, such as lighting is correlated with traffic hour and weather conditions. More refining of the independent variables might be necessary to enhance the model.

Each model has its own advantages and disadvantages. In order to produce more useful model, I need to discuss with business unit to determine what are the desirable outcome.

KNN model measures the similarity of features and uses majority rule to determine the class label, thus it is easy to understand. However, as a non-parametric model, we could not interpret which variable(s) has more effect on the target variable.

In order to evaluate each individual predictor, logistic regression model could quantitatively evaluate each variable's effect on the target variable. However, logistic regression presumes an "linear" relationship which might not stand true in this case.

Both Decision Tree and Random Forest could handle non-linearity well and could graphically illustrate the relationship between variables and target, thus they are more useful when presenting to the business units. Nonetheless, these models may be susceptible to the outliers and unbalanced distribution. When resampling, we may get a different model.

Conclusion

The modeling exercise utilized five different models, and all these models have similar performance on both training and test data set. To further improve the model performance, additional factors may be needed to determine the severity of the car accidents, such as vehicle make, weight and condition.

Additionally, the purpose of the modeling exercise may determine the preference of the model selection, whether the accuracy is more desirable or the identification of key variables is preferred. Discussion with model users would help shed lights on these areas.