Introduction

In an effort to reduce the frequency of car collisions in a community, we must create a program that will predict the severity of an accident given the current weather, road and visibility conditions. When conditions show an increased likelihood for a severe accident , this model will alert drivers to remind them to drive with more awareness and be able to alert first responders to be closer to high risk areas.

Data

The data we used is provided from coursera and has all collisions in Seattle as provided by SPD and recorded by Traffic Records from 2004 to present.
The Severity Code (‘SEVERITYCODE’) is our target variable because it is used to measure the severity of an accident either 1, less severe or 2 more severe within the dataset. We will use weather ('WEATHER'), road conditions (‘ROADCOND'), and Lighting Conditions (‘'LIGHTCOND') to train our model and be predictors for future accidents.
The data is not suitable for analysis in its given form so we will process it in order to make it usable for our objectives. We remove many of the unnecessary columns and convert features from object type to integer type. When looking at the set of data there are 3x as many type 1 in SEVERITYCODE as type 2, so we reduce the number of type 1 to create a better analysis for future events so type 1 isn’t over represented.
Now that our data has been preprocessed it is ready for analysis.

Methodology

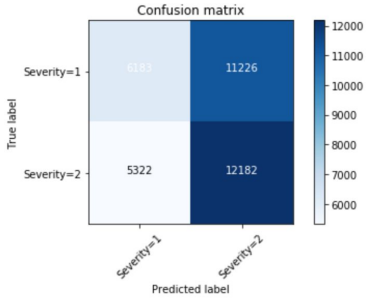
Since we are trying to predict a binary variable 1, or 2, we will process our data through Logistic Regression. This will allow us to use our historical data to predict the future outcomes in different weather, lighting, and road conditions.

We will split our data, 70% for training and 30% for training. We then build our model using our training set.

Results

We create a confusion matrix, and test the accuracy of our model using F1 score and log loss evaluation

```
Confusion matrix, without normalization
[[ 6183 11226]
 [ 5322 12182]]
```




```
In [27]: print (classification_report(y_test, yhat))
```

	precision	recall	f1-score	support
1	0.54	0.36	0.43	17409
2	0.52	0.70	0.60	17504
micro avg	0.53	0.53	0.53	34913
macro avg	0.53	0.53	0.51	34913
weighted avg	0.53	0.53	0.51	34913

```
In [28]: from sklearn.metrics import log_loss
         log_loss(y_test, yhat_prob)

Out[28]: 0.6849475402881953
```

Discussion

Using the data of car crashes from Seattle our goal was to process that data in an attempt to create a predictive model for the severity of future accidents. The data we were given had three times as many type 1 as type 2 in severity, so we downsampled to the minority so that both types had an equal number to create our model.

We assigned our weather, road conditions, and lighting conditions to a numerical value and from there we were able to create a data set that we would use to build our predictive model.

Using logistic regression we were able to train a model, and then analyze the results. Unfortunately, using the data that was provided we are unable to create an adequate predictive model for future events. Our average accuracy was 51% showing that it was not much better than a random chance

Conclusion

In conclusion we are unable to create an accurate predictive model using the data provided. There are areas that might allow for more accuracy in the future, including more specific input data, weather, road conditions, or lighting conditions. The other aspect that could help increase accuracy in our model would be having more than two severity options.