Final Report for Capstone project to predict Severity of car accident

Target readers: SPD or Traffic controlling team

Capstone Project Report

This report gives an overview of the business problem that project is going to solve and an introduction to the data set available. Also, we will define what are the steps we are going to take.

Business Problem:

The result of this study can be used by the authorities (**SPD or Traffic controlling team**) to be better prepared for accidents in the scenarios identified as critical. Also, if made publicly available people can use it to plan their drive.

We already have a data set of Seattle's car accidents. The collision data provided by SPD and traffic records from 2004 to present year.

Our goal is to find what different elements (obvious and non-obvious) resulted in an accident. Also, we will find combination of what scenarios increases the severity. A model developed with this information might help the SPD and traffic control to be well prepared in some time intervals or some weather conditions.

1. Data Understanding

The data provided by SPD and traffic Records of Seattle consists of 38 columns and around 2 lakh accidents (represented by each row). Different factors and measures for each record are put in the table format.

Now the data has several information for each accident. What kind of accident? What was the severity, Environment condition, traffic conditions, direction of cars, whether or not a pedestrian was involved etc.

Some are related to the driver, some to environmental condition and some to manual error or design of the road.

Our goal here is to find out how much role a situation played for an accident. For example, we may want to see how much role inattention plays for an accident or is it the road condition that plays important role here.

Important fields:

Some important fields from the data table are:

1. SEVERITYCODE (Text):

Severity of the accident is identified by severity code as:

• 3—fatality • 2b—serious injury • 2—injury • 1—prop damage • 0—unknown

This will also be our target variable in modelling.

2. ROADCOND (Text)

Condition of the road. (example : Wet, Dry)

3. LIGHTCOND (text)

Light conditions during the collision

4. WEATHER (text)

Description of weather during collision

5. SPEEDING (Text : Y/N)

Weather or not speeding was a factor in the collision.

6. COLLISIONTYPE (Text)

7. LOCATION (Text)

8. JUNCTIONTYPE (Text)
   Category of junction at which accident took place.

9. INATTENTIONIND (text)

   Whether or not collision was due to inattention.

10. UNDERINFL (Text)

    Was the driver under the influence of alcohol or drugs or not?

All columns from the table:

List columns to understand useful and non-useful columns

```
In [86]:  1  df.columns
```

```
Out[86]: Index(['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'],
      dtype='object')
```

After reading the data from csv, listed top 5 rows to know useful and non-useful data:

Out[85]:

	SEVERITYCODE	X	Y	OBJECTID	INCKEY	COLDETKEY	REPORTNO	STATUS	ADDRTYPE	INTKEY	...	ROADCOND	LIGHTCOND	PE
0	2	-122.323148	47.703140	1	1307	1307	3502005	Matched	Intersection	37475.0	...	Wet	Daylight	
1	1	-122.347294	47.647172	2	52200	52200	2607959	Matched	Block	NaN	...	Wet	Dark - Street Lights On	
2	1	-122.334540	47.607871	3	26700	26700	1482393	Matched	Block	NaN	...	Dry	Daylight	
3	1	-122.334803	47.604803	4	1144	1144	3503937	Matched	Block	NaN	...	Dry	Daylight	
4	2	-122.306426	47.545739	5	17700	17700	1807429	Matched	Intersection	34387.0	...	Wet	Daylight	

5 rows × 38 columns

Now we will see data-type of each column

Datatype of each column

```
In [87]: 1 df.dtypes
```

```
Out[87]: SEVERITYCODE        int64
         X                  float64
         Y                  float64
         OBJECTID             int64
         INCKEY               int64
         COLDETKEY            int64
         REPORTNO            object
         STATUS              object
         ADDRTYPE            object
         INTKEY             float64
         LOCATION            object
         EXCEPTRSNCODE       object
         EXCEPTRSNDESC       object
         SEVERITYCODE.1       int64
         SEVERITYDESC        object
         COLLISIONTYPE       object
         PERSONCOUNT          int64
         PEDCOUNT             int64
         PEDCYLCOUNT          int64
         VEHCOUNT             int64
         INCDATE             object
         INCDTTM             object
         JUNCTIONTYPE        object
         SDOT_COLCODE         int64
         SDOT_COLDESC        object
         INATTENTIONIND      object
         UNDERINFL           object
         WEATHER             object
         ROADCOND            object
         LIGHTCOND           object
         PEDROWNOTGRNT       object
         SDOTCOLNUM         float64
         SPEEDING            object
         ST_COLCODE          object
         ST_COLDESC          object
         SEGLANEKEY           int64
         CROSSWALKKEY         int64
         HITPARKEDCAR        object
         dtype: object
```

Severitycode fields study

```
In [90]: 1 df['SEVERITYCODE'].value_counts().to_frame()
```

Out[90]:

	SEVERITYCODE
1	136485
2	58188

Found only two types of severity code in database. Seems more property damages then injuries. Other severity codes given in metadata but not present in database.

Once we have analysed the data columns, rows, target field, we can start on data preparation and cleaning.

3. Methodology:

For Jupyter notebook of below anyalysis please refer below link:

https://github.com/ejly/Capstone-Project---Car-accident-severity

Data preparation and cleaning:

For gathering the data, we will use Pandas library in Jupyter notebook. For preparing the data for analysis, we need to drop irrelevant data columns. Fix data types and formats if required.

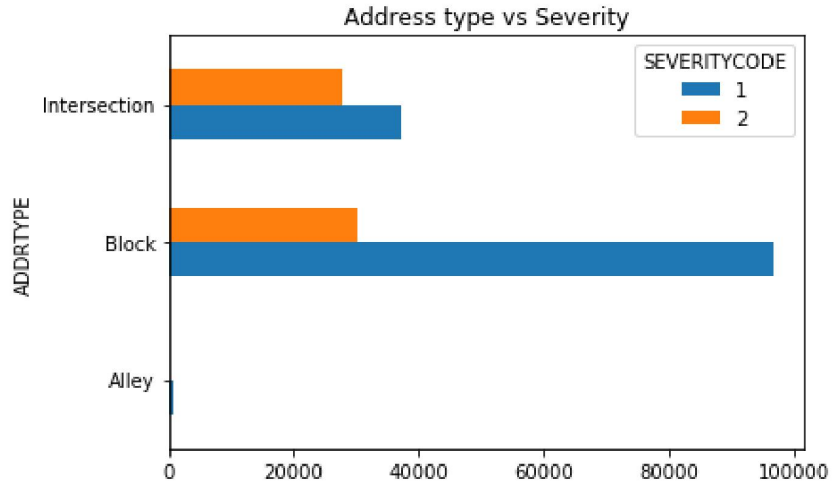
Some columns that are not relevant were dropped:

Dropping few irrelevant fields from begining

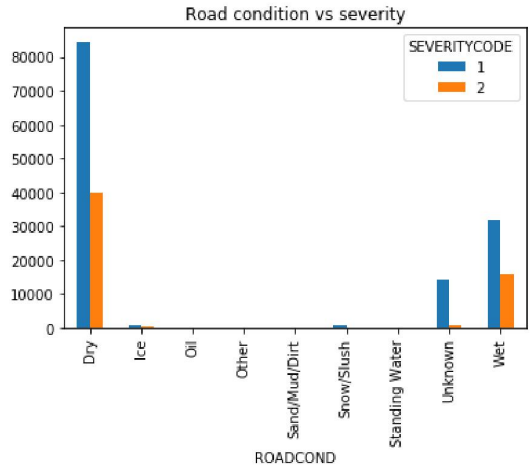
```
In [88]:  1  df.drop(columns=['X', 'Y', 'OBJECTID','INCKEY', 'COLDETKEY'], inplace=True)

In [89]:  1  df.columns

Out[89]: Index(['SEVERITYCODE', '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'],
               dtype='object')
```

Many cells that didn't have any value were also replaced with "Unknown" string or "Others" string.

For gathering information as how much if any impact of individual column has on severity, we plotted bar graphs. These all include the categorial data v/s the Severity code.



Finding: There are certain places that gets more collision (this is obvious)



Finding: There are many accidents that happened on dry roads. Though Wet roads also caused around 1/3rd of the total

Similarly, all below columns seem to have impact on severity code:

ADDRTYPE', 'COLLISIONTYPE', 'JUNCTIONTYPE', 'SDOT_COLDESC', 'INATTENTIONIND', 'UNDERINFL', 'WEATHER', 'ROADCOND', 'LIGHTCOND', 'SPEEDING', 'ST_COLDESC', 'HITPARKEDCAR'

From above we found our relevant columns, and these will be used for creating our decision tree model.

So, we select “SEVERITYCODE” as our prediction column (y)

And relevant columns were transformed to numerical values.

Using the scikit we got train and test data. Then predict yhat to know the precision achived.

This we calculated through F1-Score as given in next section

4. Results:

With Decision tree model we were able to create a model with good precision and thus can be used with confidence. (F1 score : 0.823)

In [114]:
```
X_train
y_train
tree.fit(X_train,y_train)
yhat = tree.predict(X_test)

print("F1 score: ", metrics.f1_score(y_test, yhat, labels=[1]))
```

```
F1 score:  0.8237429552083838
```

5. Conclusion:

With the below variables we can predict the severity with high confidence.

"ADDRTYPE', 'COLLISIONTYPE', 'JUNCTIONTYPE', 'SDOT_COLDESC', 'INATTENTIONIND', 'UNDERINFL', 'WEATHER', 'ROADCOND', 'LIGHTCOND', 'SPEEDING', 'ST_COLDESC', 'HITPARKEDCAR'

These attributes can be selected, and an application can be created for Traffic control team to better be prepared on the severity of accident in a locality depending on Junction, weather/road/light condition.

Also, data can be made public to alert Inattention and “driving under influence of alcohol and drugs” can increase the severity of an accident.