**Predicting the Severity of Accidents in Seattle**

**Shwetha Acharya**

**Oct 22, 2020**

1. **Introduction**
   1. **Background**

**In the recent past we have seen an increase in the number of road accidents in Seattle. While some of these accidents just cause property damage some are serious, resulting in heavy injuries and fatalities to drivers and passengers and pedestrians. With this trend, driving on highways and roads poses a risk to the lives of people. There may be different reasons for these accidents such as poor road conditions, light and weather conditions and so on. The idea is to use data science techniques to understand if we can predict the gravity of the accidents on account of collisions.**

* 1. **Problem**

**Data that might contribute to determining accident severity might include road conditions, light and weather conditions, car speeding, drunken driving. This project aims at predicting accident severity based on historic data.**

* 1. **Interest**

**This analysis and prediction will help drivers and passengers of various vehicles such as trucks, cars as well as pedestrians taking these road routes as it will serve as a safety warning to them. In addition, it would help traffic cops monitoring these areas – making them more vigilant and taking steps to avoid collisions and accidents, thus saving/securing lives of people.**

1. **Data Acquisition**
   1. **Data Sources**

**Data for this study is taken from Seattle Department of Transportation - Traffic Management Division, Traffic Records Group\*\*. It includes information on all types of collisions from 2004 till date (2020). The key features in the dataset encompass accident severity in terms of numeric codes, collision description, address type (whether intersection, alley or block), X and Y co-ordinates, number of vehicles and pedestrians involved, light, road and weather conditions, speeding, whether driving under the influence**.

* 1. **Data Cleaning**

**Our target variable is the accident " severity" in terms of human fatality, traffic delay, property damage and hence SEVERITYCODE is the target variable – y. There are 37 features - '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'**

**On observing, we see that following columns can be dropped for the below reasons**:

* **EXCEPTRSNCODE, EXCEPTRSNDESC – Metadata doesn’t explain much about these columns. Also, columns have many null values and do not have any impact on the prediction.**
* **SEVERITYCODE.1 is a duplicate column**
* **OBJECTID and REPORTNO are just unique identifiers or keys to identify every row and collision report and no impact on accident predictions.**
* **SEVERITYDESC is just the accident description - Numerical values are captured in SEVERITYCODE.**
* **INCDATE - It is a subset of INCDTTM**
* **INCKEY and COLDETKEY – These are unique keys for the incident to identify collision**
* **On inspecting SDOT\_COLCODE, SDOT\_COLDESC, ST\_COLCODE, ST\_COLDESC**

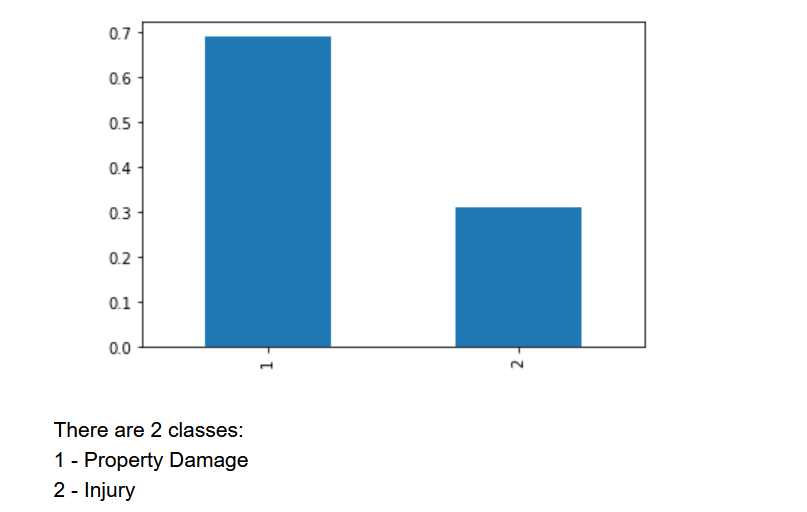
1. **ST\_COLCODE and ST\_COLDESC -refer to code provided by the state that describes the collision**
2. **SDOT\_COLCODE and SDOT\_COLDESC – refer to code given to the collision by SDOT.**

* **One of the above can be dropped as it is a code describing the same collision but given by 2 authorities - SDOT/State. Given that, ST\_COLCODE has some null values when compared to SDOT\_COLCODE, ST\_COLCODE and ST\_COLDESC can be dropped. Also, SDOT\_COLDESC - description of SDOT\_COLCODE, can be dropped**
* **LOCATION – Location/address is given with X and Y co-ordinates/latitude-longitude and hence location can be dropped**
* **Also, more than 95% of data is missing for ‘SPEEDING’ and ‘PEDESTRIAN-RIGHT-OF-WAY-NOT-GRANTED’ and hence they can be dropped**
* **INTKEY can be dropped too, as it is a key to represent address type - “Intersection” of a given location.**

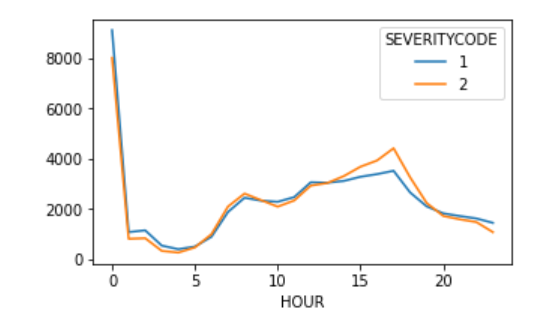
**X and Y represent latitudes and longitudes of the accident sites. They can be used in our model development after normalization. Categorical variables such as Address Type, Junction Type, Weather, Road and Light condition are converted to numeric values. INCDTTM is a date column and fine grained to classify into weekend and non-weekend days.**

**After cleaning up this data, the above set of refined attributes is put through feature selection to pick only the relevant features.**

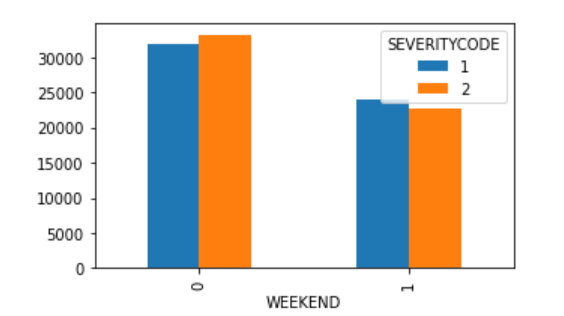
**Upon examining the distribution of classes for severity code, we see that data is imbalanced.**



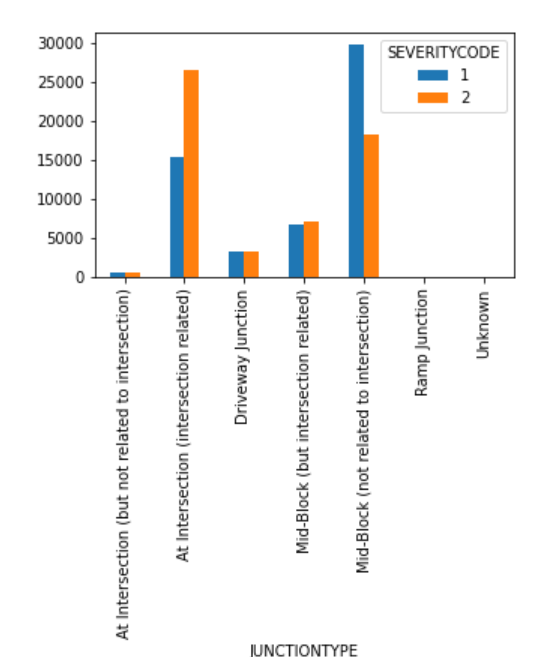
**Model performance is affected by imbalance in the classes so we will under-sample the dominant class to rebalance the dataset. The resulting dataset will be smaller, but this will actually make the computations faster.**

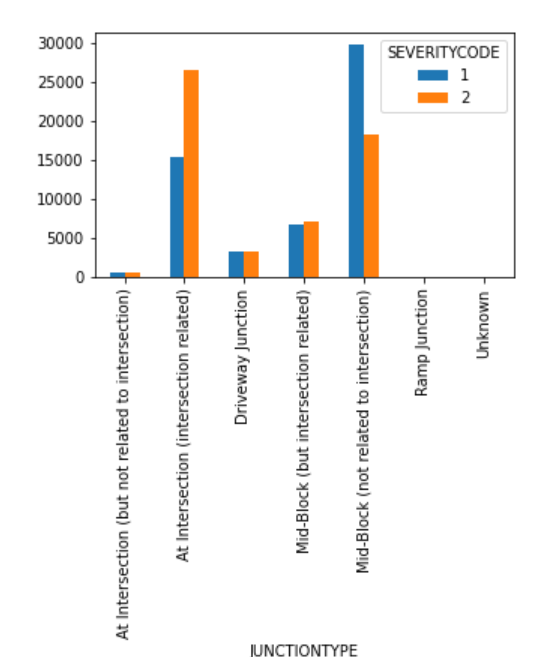


**Examining the relationship between HOUR attribute and SEVERITY, I understand that a**t midnight and in the afternoon/early evening, number of accident cases is on the higher side. At early evening, injury causing accidents are more.

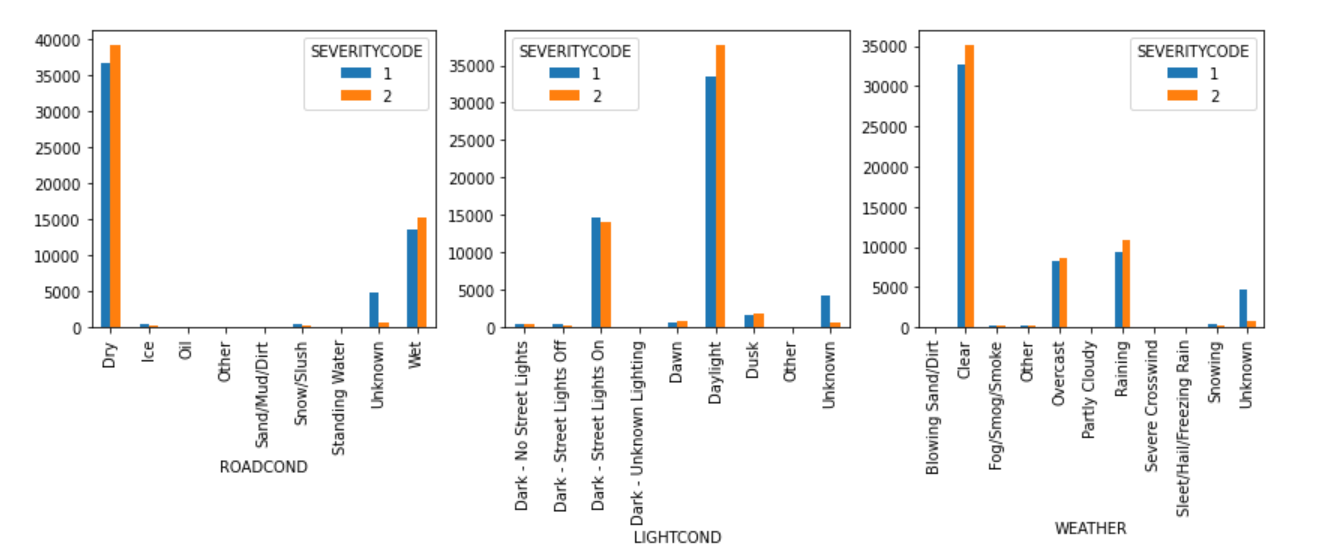


**Also, on Weekends there are comparatively less number of accidents. Out of total number of accidents, there is an equal proportion of property damage and injury-based accidents be it weekends or weekday**s.



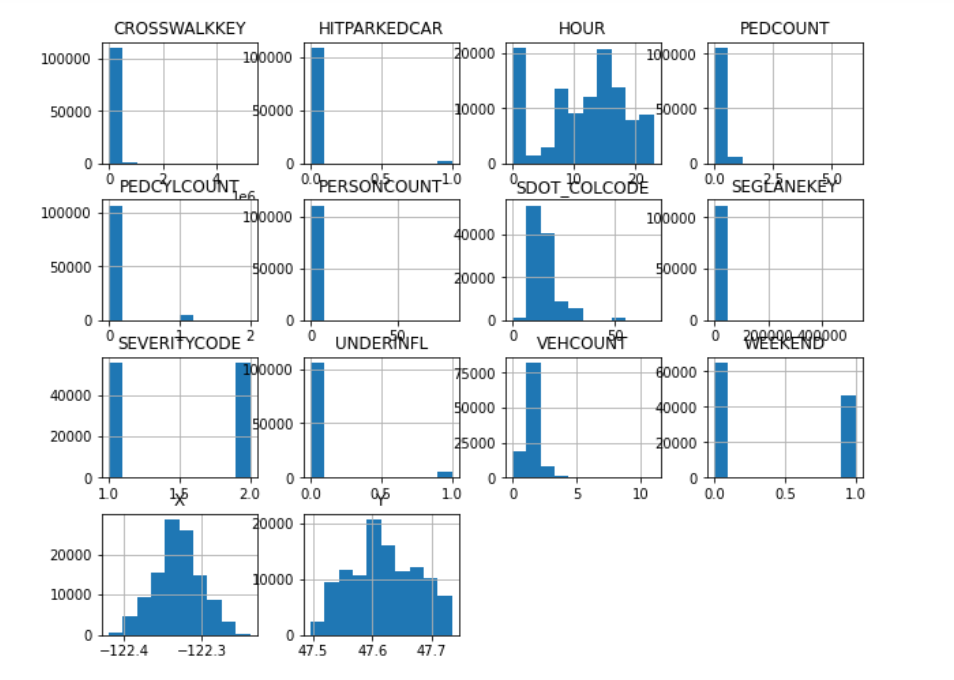


Property damage is maximum with accidents at “Midblock”, while “at intersection” injury resulting accidents are maximum .

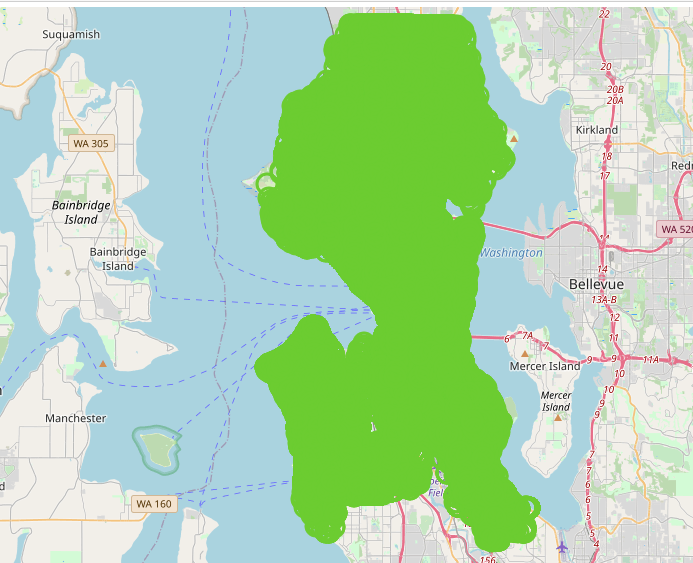


**Majority of accidents seem to happen when roads are dry in daylight and clear weather**

**Frequency distribution of the numeric attributes shows that most of the variables are either constant(with little variance) or skewed**

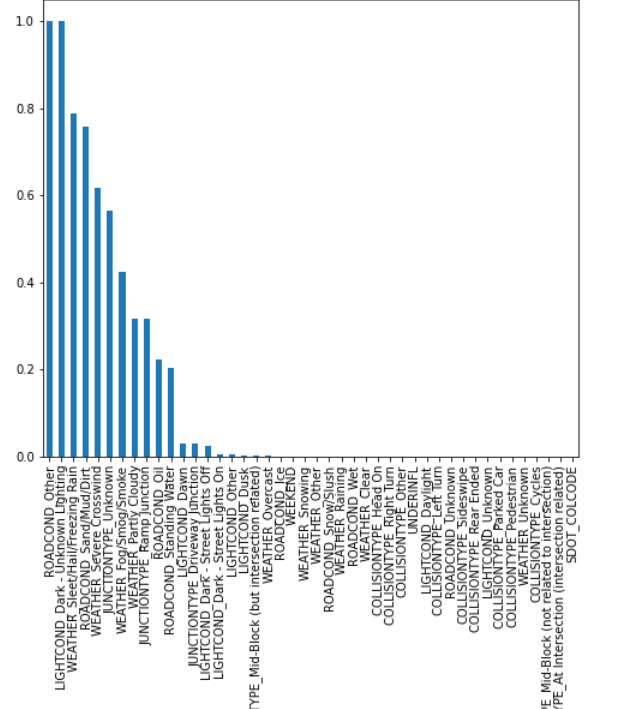


**On plotting X and Y co-ordinates on a map, I see that accident locations are spread across Seattle. Almost all areas are covered.**



* 1. **Feature Selection**

**After converting categorical variables to numeric, we have** 55 attributes now and we need to perform some feature selection to pick the ones that influence the severity code the most. We will perform chi squared test on the categorical variables and get the p-values. Features with p-value > 0.05 can be rejected because greater p-value indicates no relationship between the target and itself.



Based on chi squared test results, I rejected features with p-value > 0.05. Also, 'SEGLANEKEY', 'CROSSWALKKEY', 'HITPARKEDCAR' are almost constant for the whole dataset. These features do not show any patterns of variation.

**The Final 40 attributes for model development include**

**'ROADCOND\_Ice', 'ROADCOND\_Snow/Slush', 'ROADCOND\_Unknown', 'ROADCOND\_Wet',**

**'WEATHER\_Clear', 'WEATHER\_Other',**

**'WEATHER\_Overcast', 'WEATHER\_Raining',**

**'WEATHER\_Sleet/Hail/Freezing Rain',**

**'WEATHER\_Snowing', 'WEATHER\_Unknown',**

**'LIGHTCOND\_Dark - Street Lights Off',**

**'LIGHTCOND\_Dark - Street Lights On',**

**'LIGHTCOND\_Dawn',**

**'LIGHTCOND\_Daylight', 'LIGHTCOND\_Dusk', 'LIGHTCOND\_Other',**

**'LIGHTCOND\_Unknown',**

**'JUNCTIONTYPE\_At Intersection (intersection related)',**

**'JUNCTIONTYPE\_Mid-Block (but intersection related)',**

**'JUNCTIONTYPE\_Mid-Block (not related to intersection)',**

**'COLLISIONTYPE\_Cycles', 'COLLISIONTYPE\_Head On',**

**'COLLISIONTYPE\_Left Turn', 'COLLISIONTYPE\_Other',**

**'COLLISIONTYPE\_Parked Car', 'COLLISIONTYPE\_Pedestrian',**

**'COLLISIONTYPE\_Rear Ended', 'COLLISIONTYPE\_Right Turn',**

**'COLLISIONTYPE\_Sideswipe', 'X', 'Y', 'PERSONCOUNT', 'PEDCOUNT',**

**'PEDCYLCOUNT', 'VEHCOUNT', 'SDOT\_COLCODE', 'UNDERINFL', 'HOUR', 'WEEKEND'**

**This dataset is normalized using standard scalar and split into train and test sets.**

1. **Modeling and Evaluation**

**We have used classification algorithms to predict severity codes. Following algorithms were used:**

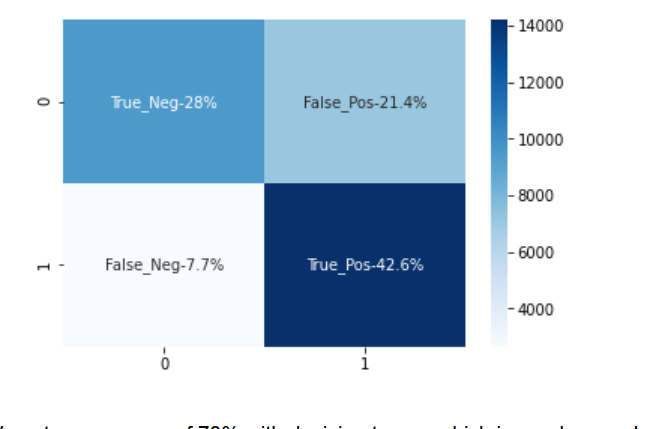
1. **Decision Trees**
2. **RandomForest Classifier**
3. **SVM**
4. **Logistic Regression**
5. **XGBoost RF Classifier**

**Accuracy\_score, f1 score , confusion matrix and cross validation scores were used to evaluate models and models exhibited an accuracy of around 70+% with Decision trees, XGBOOST, RandomForest Trees and logistic regression with slight variation.**

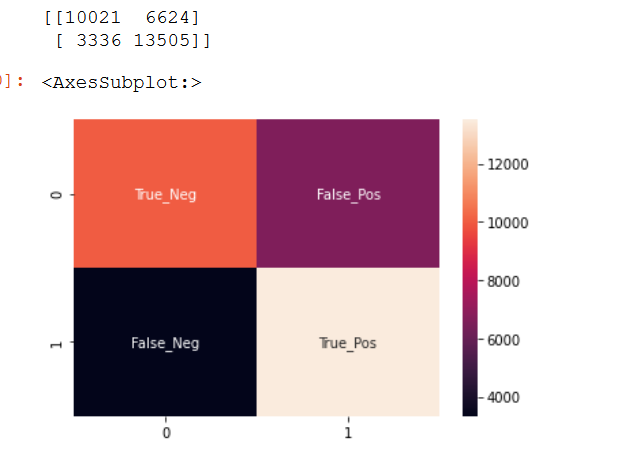
**However, SVM takes 80-90% more time than other models to execute.**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Classification Model** | **Accuracy** | **F1-Score** | **Cross Val Score** | **Log Loss** |
| **DecisionTrees** | **0.706** | **0.699** | **0.6998** | **NA** |
| **RandomForest** | **0.71** | **0.678** | **0.7** | **NA** |
| **SVM** | **0.7** | **0.698** |  | **NA** |
| **Logistic Regression** | **0.702** | **0.698** | **0.7** | **0.54** |
| **XGBoost Random Forest Classifier** | **0.7146** | **0.706** | **0.7** | **NA** |

**Decision Trees confusion Matrix**



**Logistic Regression Confusion Matrix- Percentage of false positives is on the higher side**



**Though scores are above average, there is room for improvement with these models.**

### Conclusion

**In this study, we analyzed the relationship between accident severity and general environment conditions and the road lanes. We identified that some of the weather, road, light conditions, are among the most important features that affect the severity. Classification models were built to predict the severity of accidents and this information can be very useful to people taking the corresponding routes and to the traffic department. Model accuracy can be improved by getting data on other features such as attention indicators and speeding, driver’s profile, vehicle working status or any related events which may lead to distraction/deviation.**

**\*\*Data used for this analysis is taken from coursera courtesy SDOT Traffic Dept**