Predicting the Severity of an Accident due to Environmental Aspects and Causes

Freddie Perez

October 13th , 2020

Business Problem

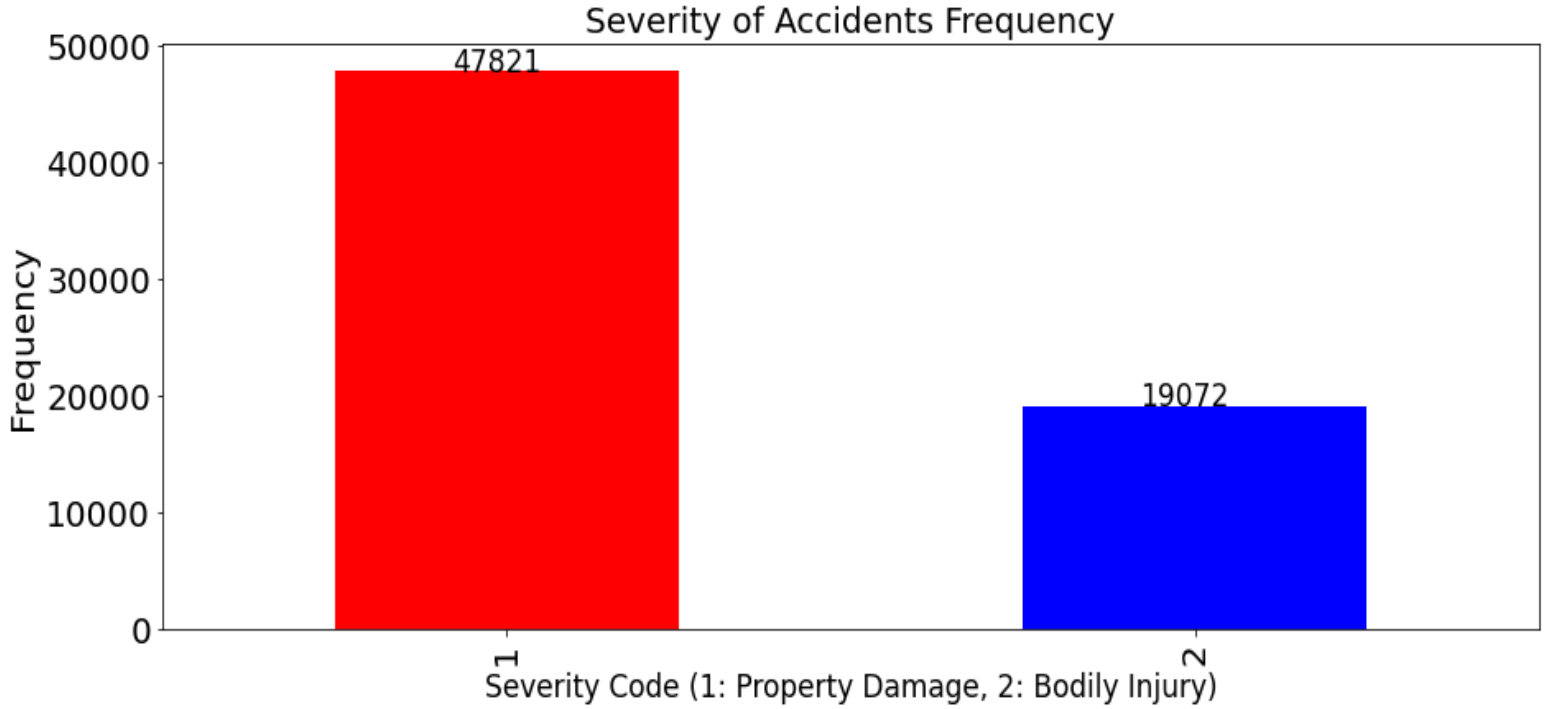
- Is it possible to predict the severity of an accident due to various environment information?
- If we are able to find a solution for this, we can help potential stakeholders, everyday users of vehicles in this case, by providing insight on the risks associated with traveling.
- These insights on the risks can then allow the stakeholders to Help guide their transportation decisions and actions such as:
  - Choosing a different method of transportation
  - Being more attentive and adhering to defensive driving maneuvers

Data Acquisitions and Cleaning

- The dataset I used was provided by the course and involves collision data from across Seattle, Washington.
- In total, it has upwards of 70,000 observations, with 37 potential attributes of interests in the raw dataset.
  - Any non-informative attributes, and those that do not pertain to environmental aspects, were dropped.
- Attributes kept includes the target attribute ‘SEVERITYCODE’, as well as ‘WEATHER’, ‘ROADCOND’, ‘LIGHTCOND’.
- After cleaning and data munging has been completed, 66893 observations were left.
- 25 attributes that consisted of One-Hot Encoding on categorical attributes.

Imbalanced Dataset

- After data cleaning, I plotted the distribution of the observations with respect to the SEVERITYCODE attribute which yielded:



Imbalanced Dataset

- In order to counteract the imbalance issue between the two classes, I did the following to the cleaned dataset:
  - Created a training and testing data set that were stratified such that the proportions matched the original dataset.
  - Created another set of training and testing datasets where I applied Oversampling to the Minority Class (SEVERITYCODE 2: Bodily Injury)
  - Created a third set of training and testing datasets where I applied the SMOTE Technique (Synthetic Minority Oversampling Technique) in conjunction with Under sampling the Majority class.

Machine Learning Models Used

- The machine learning models I used for this experiment consisted of the following:
  - **K-Nearest Neighbors (with K=2, as it had the highest Accuracy)**
  - **Decision Trees**
  - **Random Forest Ensemble Method**
  - **Support Vector Machines**
  - **Logistic Regression**
- Prior to applying these techniques, I normalized that data as well for the techniques that rely on the distance metric.
- I also tested all three types of datasets with each Machine Learning technique (Stratified, Oversampled Minority Class, SMOTE/Undersampled Majority Class)

Machine Learning Models Evaluations

- For evaluating the machine learning techniques, I used the Average F1 Score, Jaccard Similarity Score, as well as the LogLoss Score (for the Logistic Regression) metrics.
  - Using these metrics should provide more insight given the imbalanced characteristic of the initial raw data.

KNN	Jaccard	F1-score	LogLoss
Stratified	0.6386	0.6020	NA
Oversampled	0.7126	0.5965	NA
SMOTE/Undersampled	0.7146	0.5965	NA

Decision Tree	Jaccard	F1-score	LogLoss
Stratified	0.6508	0.6262	NA
Oversampled	0.6522	0.6254	NA
SMOTE/Undersampled	0.3047	0.4632	NA

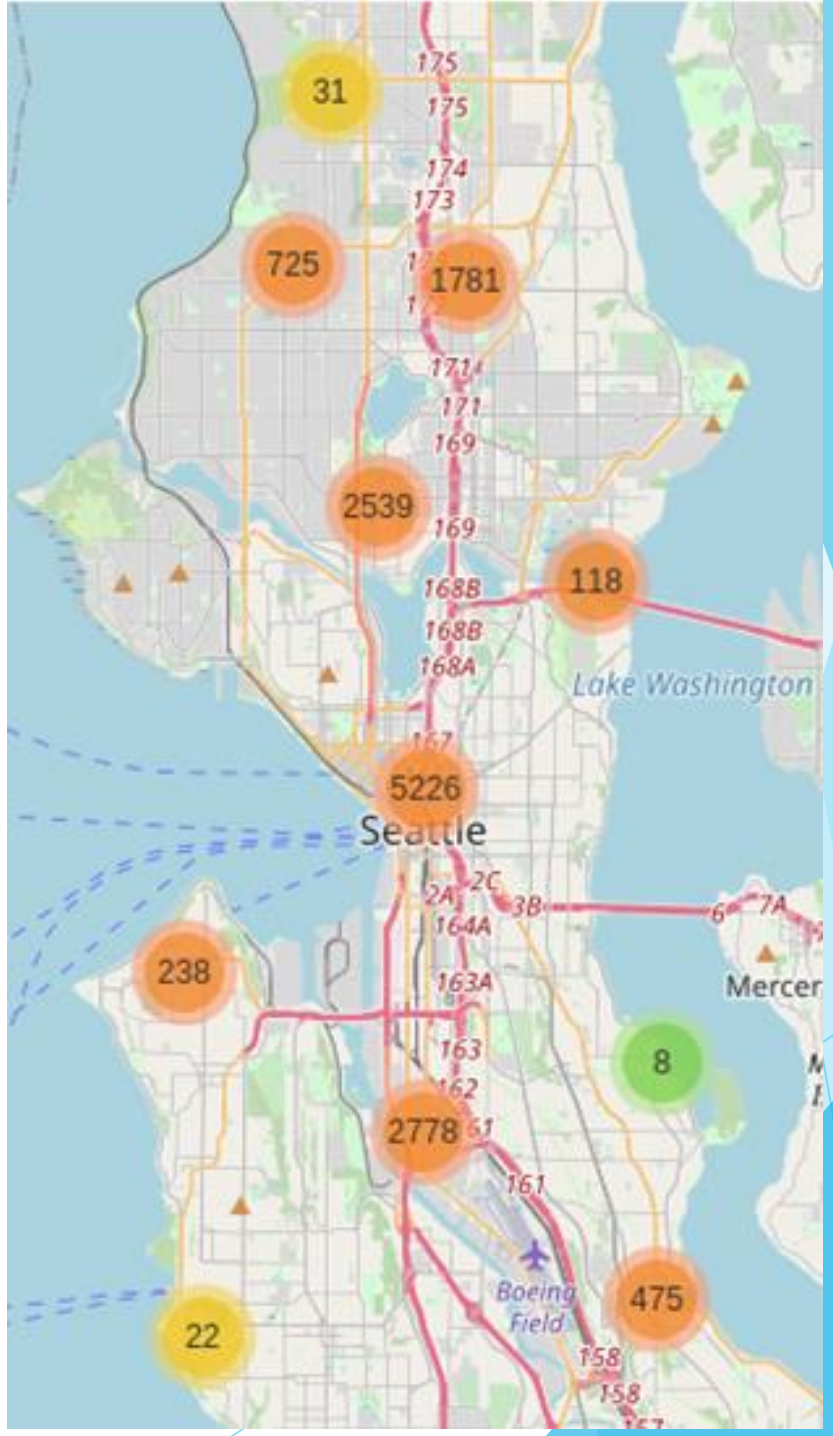
Random Forest	Jaccard	F1-score	LogLoss
Stratified	0.7148	0.5967	NA
Oversampled	0.3232	0.4773	NA
SMOTE/Undersampled	0.3034	0.4632	NA

SVM	Jaccard	F1-score	LogLoss
Stratified	0.7148	0.5964	NA
Oversampled	0.3220	0.4764	NA
SMOTE/Undersampled	0.3036	0.4627	NA

Logistic Regression	Jaccard	F1-score	LogLoss
Stratified	0.7149	0.5960	0.58
Oversampled	0.3567	0.5011	0.66
SMOTE/Undersampled	0.2793	0.4424	0.68

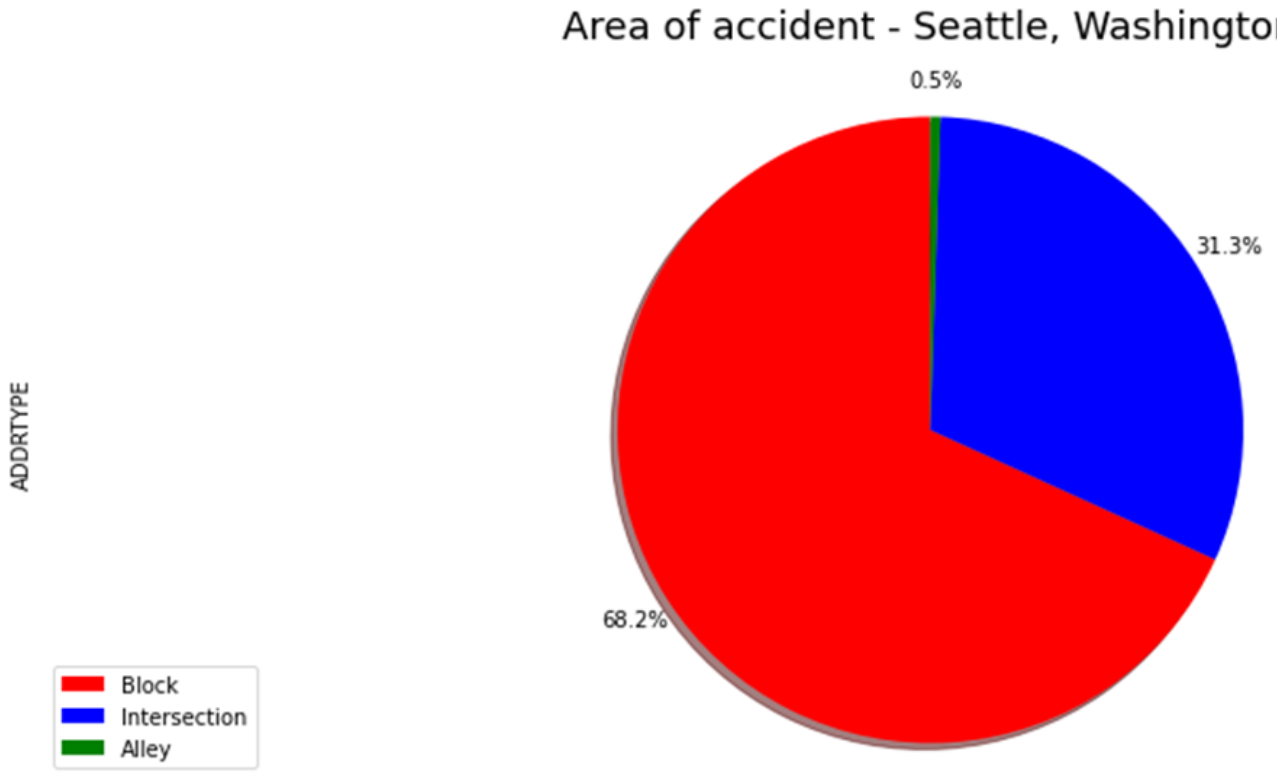
Recommendations

- Through the exploration process of the data and the machine learning model building process, various insights were brought up that could be of use for the stakeholders at hand.
- One insight is with respect to the locations that had the highest frequency of accidents, which were primarily centered on interstate-5 as it passes through highly dense, populated areas.



Recommendations

- We can also see that the 68.2% of accidents occur around the city blocks, with intersection being a close second at 31.3%.
- In addition, the majority of accidents were caused by either severe weather, poor lighting conditions, poor driving conditions, or a combination of all 3 major attributes.



Recommendations

- In addition, the majority of accidents were caused by either severe weather, poor lighting conditions, poor driving conditions, or a combination of all 3 major attributes.
- In general, methods to mitigate the potential to incur the risk of having some type of accident occur on your personal property would be to:
  - Note the flow of traffic
  - Increase your following distance to allow for better breaking opportunity
  - As well driving more defensively and safely during adverse conditions

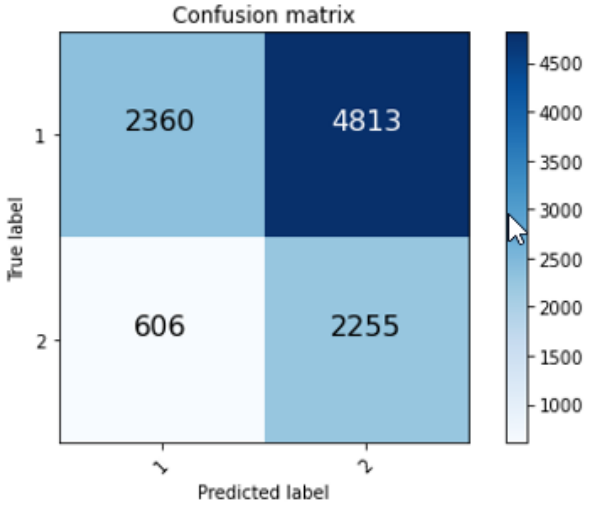
Conclusion

- Overall, it seemed like each model did average in terms of performance. None of them performed excellently or extraordinarily.
  - The accuracy was always capped at 0.71 or 71%.
- Even with the use of Average F1 Scores as an additional metric, looking Precision and Recall showed that most of the models were always skewed toward one class or the other (Severity Code 1 or 2).
  - This could primarily be due to having a very skewed/imbalanced dataset to begin work with.
- The classification report results from the Random Forest model, SVM model, and the Logistic Regression, all with the dataset that had `SMOTE/Under sampling` applied, had a very high precision score of 0.8 for class 1 and recall value of 0.79-0.82 for class 2.
- Meanwhile, the Recall was very low for class 1, and the Precision was also low for class two, despite any hyper-tuning applied to the models.

Conclusion (Model Stats)

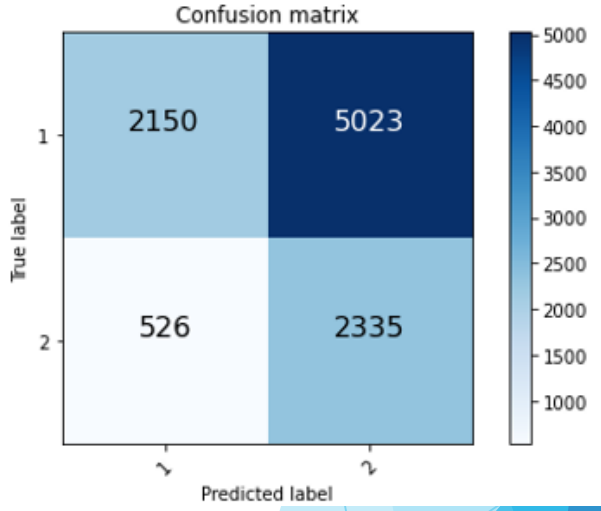
Random Forest (SMOTE/Under Sampling)

	Precision	Recall	F1-Score
1	0.80	0.33	0.47
2	0.32	0.79	0.45
Macro Avg.	0.56	0.56	0.46
Weighted Avg.	0.66	0.46	0.46



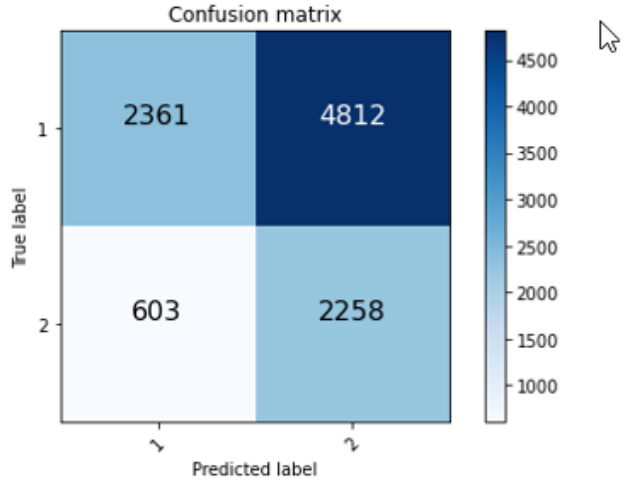
Logistic Regression (SMOTE/Under Sampling)

	Precision	Recall	F1-Score
1	0.80	0.30	0.47
2	0.32	0.82	0.45
Macro Avg.	0.56	0.56	0.46
Weighted Avg.	0.66	0.46	0.46



Support Vector Machine (SMOTE/Under Sampling)

	Precision	Recall	F1-Score
1	0.80	0.33	0.47
2	0.32	0.79	0.45
Macro Avg.	0.56	0.56	0.46
Weighted Avg.	0.66	0.46	0.46



Conclusion and Future Steps

- Initially, I expected the application of the sampling techniques would help the models perform better.
- The Stratified Datasets, which created training and testing data sets based on the same proportions of each class in the clean data, performed much better.
- Methods of building an improved model would include:
  - Using a more balanced Dataset
  - Having more than 68,000 observations to properly training the Machine Learning Algorithms
- If more time was available, the application of Time-Series modeling could be the next step.
  - This would add a temporal overlay that can add temporal based attributes such as time, date, and month as part of the learning process.
  - This may further improve predictive power the model if it comes to fruition.