1.INTRODUCTION TO BUSINESS PROBLEM

The world as a whole suffers due to car accidents. According to the Global status report on road safety 2018, launched by WHO in December 2018, highlights that the number of annual road traffic deaths has reached 1.35 million. Road traffic injuries are now the leading killer of people aged 5-29 years.

The 2020 report of the National Highway Traffic Safety Administration suggests that USA is no exception compared to the rest of the world. According to the report, motor vehicle crashes are the number one safety problem in American transportation. They account for 94 percent of transportation death and 99 percent of transportation injury.

This project aims to analyze the relation of several factors and their impact to the severity of accidents, to predict how severity of accidents can be reduced based on these factors, and provide insights to the possible stakeholders such as car drivers and Public Development Authorities.

2.DATA

The dataset used for this project is based on car accidents which have taken place within the city of Seattle, Washington from the year 2004 to 2020. This data is regarding the severity of each car accidents along with the time and conditions under which each accident occurred. The model aims to predict the severity of an accident, considering that, the variable of Severity Code was in the form of 1 (Property Damage Only) and 2 (Physical Injury).

The dataset includes several variables that could directly or indirectly affect the severity of accidents. The variables are analyzed to check if their effect is statistically significant against the target variable 'severity'. The variables that are chosen to be used in the development of the model are processed the deal with missing values and form the final version of the data to be trained.

After examining the meta-data document the features that are directly related to the target variable ‘SEVERITY_CODE’ such as ‘SEVERITY_DESC’, ‘ INCKEY’, ‘OBJECTID’ are dropped. Then

using a heat map and inspecting correlations, variables 'SPEEDING', 'PEDCOUNT', 'PEDCYLCOUNT', 'PEDROWNOTGRNT','VEHCOUNT', 'INATTENTIONIND', 'UNDERINFL', 'HITPARKEDCAR', 'MONTH', 'WEATHER', 'ROADCOND', 'LIGHTCOND', 'ADDRTYPE', 'COLLISIONTYPE', 'TIME' are selected as predictor variables.

3.DATA CLEANING AND PREPROCESSING

Considering that generally traffic accidents depends on the time of the day and month of the year by processing the 'INCDTTM' attribute new features 'MONTH' and 'TIME' are added to the dataframe.

```
dataframe['INCDTTM'] = pd.to_datetime(dataframe['INCDTTM'])
dataframe['MONTH'] = dataframe['INCDTTM'].dt.month
dataframe['TIME'] = dataframe['INCDTTM'].dt.time
dataframe['TIME'] = pd.to_datetime(dataframe['INCDTTM'])
```

```
dataframe.drop(columns='INCDTTM',inplace=True)
```

```
dataframe['TIME'] = (dataframe['TIME'].dt.hour % 24 + 4) // 4
dataframe['TIME'].replace({1: 'Late Night',
                           2: 'Early Morning',
                           3: 'Morning',
                           4: 'Noon',
                           5: 'Evening',
                           6: 'Night'}, inplace=True)
```

Figure 1: 'INCDTTM' Attribute Processing

After inspecting the missing values the variables that have a very large missing value count are dropped namely 'SPEEDING' and 'INATTENTIONIND' columns dropped completely, and for the remaining the rows containing missing values are dropped since they were few enough to discard compared to the number of total observations in the dataset.

```
df.isna().sum()
```

```
X                 5334
Y                 5334
SEVERITYCODE         0
WEATHER           5081
ROADCOND          5012
LIGHTCOND         5170
SPEEDING        185340
ADDRTYPE          1926
PEDCOUNT             0
PEDCYLCOUNT          0
PEDROWNOTGRNT        0
VEHCOUNT             0
INATTENTIONIND  164868
UNDERINFL         4884
HITPARKEDCAR         0
COLLISIONTYPE     4904
MONTH                0
TIME                 0
dtype: int64
```

Figure2: Missing Values

Since neither the Seaborn's heatmap nor the mutual information feature selection proved any set of variables are distinguishably better than others, and since the number of observations are enough not to discard any features, all features other than the ones that are directly tied to SEVERITYCODE or the ones with a very large number of missing values are kept as predictor variables.

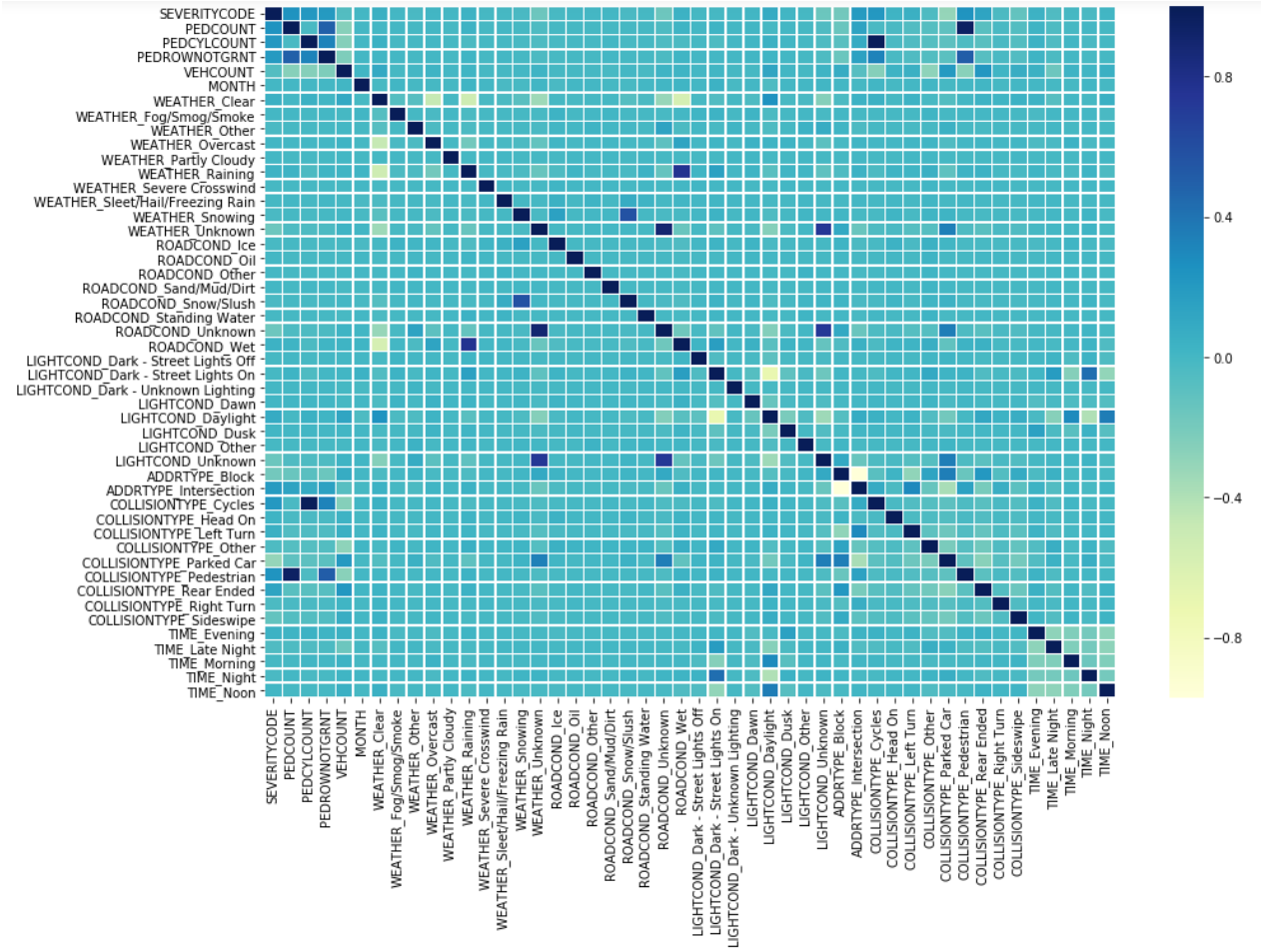

Figure 3: Heatmap

For the categorical variables such as WEATHER, ROADCOND, COLLISIONTYPE etc., one hot encoding is used instead of label encoder in order not to introduce any non-existing ordinality in the features.

Finally, in order to compensate for the imbalance in the dataset where there is a 7:3 ratio between the majority class of property damage and the class of injury. Random under sampling is used to balance the data on the training data set instead of SMOTE or over sampling since the dataset is very large and under sampling saves computing time and shown to perform better than SMOTE with certain machine learning models e.g. XGBoost or Random Forest.[1]

4.MODEL SELECTION

The following models were selected: XGBoost, RandomForest, KNN and SVM, to fit on the training set. The model's F1 score was then calculated on the test set. The table in the Results section shows their performance.

RESULTS

	MODEL	JACCARD SCORE	F1_SCORE
0	KNN	0.3929	0.7004
1	SVM	0.4259	0.6794
2	XGBoost	0.4259	0.6794
3	Random Forest	0.3988	0.6685

Based on the performance above KNN is the the best performing model.

5. CONCLUSION

The goal of this project is to analyze historical vehicle crash data and classify the severity of an accident using environmental condition and accident features weather, road, time of the day, time of the year etc. Vehicle accident data from the City of Seattle's' Police Department for the years 2004 until present were used. The data was cleaned, and features related to environmental conditions were selected and analyzed. It was found that majority of accidents happened in clear weather, dry roads, and during daytime which was intuitively unexpected. Machine learning models: K-Nearest Neighbor, SVM, XGBoost and Random Forest were used to predict the severity of an accident based on certain environmental conditions. The models used were also evaluated using different accuracy metrics that can be used regardless of the imbalance of classes.

6.DISCUSSION

Given the findings of exploratory analysis:

The authorities could:

Introduce more control in traffic during noon and within blocks, can increase the penalty for not granting pedestrian right.

Collect additional data on accidents such as vehicle speed on the time of accident, or visuals captured by traffic cameras that could help improve the classification model in the future.

6. REFERENCES

[1] Handling Imbalanced Data: SMOTE vs. Random Under sampling, S Mishra - Int. Res. J. Eng. Technol, 2017 - academia.edu