U.S. Road Accidents Severity Prediction

Team 52 – Final Report

Ashish Singh (903650462), Hari Rishi Bahadur (903744824),

Mujtaba Hussain Mohammed (903737808) and Vivek Sagi (903045070)

Problem Statement

Our group's problem statement is as follows.

- How is the severity of US road accident linked to
    a. Prevalent weather conditions at the time of the incident (Primary Research question)
    b. Socio-economic factor i.e., median income of the population in a zip code (Secondary Research question 1)
    c. Lockdowns / stay at home advisory during Covid (Secondary Research question 2)

Motivation (Choice of Topic)

The economic and societal impact of traffic accidents cost U.S. citizens hundreds of billions of dollars every year. A large part of losses is caused by a small number of serious accidents. Reducing traffic accidents, especially serious accidents, is nevertheless always an important challenge. The proactive approach, one of the two main approaches for dealing with this problem, focuses on preventing potential unsafe road conditions. For the effectively implementing this approach, accident prediction and severity prediction is critical. If we can identify the patterns of how these serious accidents happen and the key factors, we may be able to implement well-informed actions and better allocate resources.

The primary intent behind selecting the chosen project topic is to identify if there exists a correlation between the road accidents in the US and the prevalent weather conditions during the incident. Some examples of weather-related variables are visibility, wind direction, wind speed, weather conditions – overcast, light rain, heavy rain. In addition, we also want to analyse if the US road accident severity is also impacted by the following 2 (two) aspects.

- Covid related lockdowns / stay at home advisories.
- Socio-economic conditions i.e., median income of a population aligned to a zip code.

Business Justification

We think such research will be beneficial for an insurance company. It should positively impact all the following aspects related to their line of business as such an analysis will assist them in understanding the factors which may contribute to the severity of accidents and in turn effect the claim figures.

- Financial: A richer understanding of the factor contributing to severe accidents will allow the insurance companies to identify if any of those can be controlled. If yes, then by effectively controlling such factors will lead to less severe accidents and in turn lower claim figures. This will then directly impact the bottom line of the companies in a positive manner.
- Marketing: Based on the knowledge of the factors that contribute towards severe accidents, the insurance companies can roll out offerings which will instil safe driving behaviours. An example of this is the Drive Safe app from All State insurance. The company incentivizes customers based on how they drive and the time of the day when they drive.
- Operational: Once they have a fair forecast of the claim rates, it can help them achieve operational efficiency by sizing their operations team to process claims, engaging an adequate

number of inspectors for site inspection and staff the customer service team so that appropriate levels of customer service can be provisioned.

Initial Hypothesis

Our initial hypothesis for the 3 (three) research questions (1 (one) primary and 2 (two) secondary) is as follows:

- Primary Research Question: Weather related parameters have a correlation on the severity of an accident.
- Secondary Research Question 1: Low-income neighbourhood have more severe accidents.
- Secondary Research Question 2: Accident severity did not change between the pre-Covid and the Covid time periods.

Understanding of the Data

Datasets

To sufficiently analyse the primary and the secondary research questions, three datasets are used.

- **Accident Dataset:** This is a countrywide car accident dataset, which covers 49 (forty-nine) states of the USA. The accident data is collected from February 2016 to Dec 2021, using multiple APIs that provide streaming traffic incident (or event) data. These APIs broadcast traffic data captured by a variety of entities, such as the US and state departments of transportation, law enforcement agencies, traffic cameras, and traffic sensors within the road-networks. Currently, there are about 2.8 million accident records in this dataset.
- **Income Dataset:** The dataset filters on income and related statistics made available through the 2011-2015 ACS 5-Year Documentation was provided by the U.S. Census Reports. Income is a vital element when determining both quality and socioeconomic features of a given geographic location. The dataset has +32,000 records, with granularity on a neighbourhood scale (mean, median, standard deviation). The data was derived from over +36,000 files and covers 348,893 location records.
- **Inflation Dataset:** It contains the inflation data from 2016 through 2021 for all items less food and energy in U.S. city average, all urban consumers, not seasonally adjusted. This dataset is sourced from U.S. Bureau of Labor Statistics. This dataset is used only to adjust the income in the Income dataset for inflation.

Key attributes for the Accident and Income datasets are analyzed in the project proposal.

Data Wrangling

Accident Dataset

- **Dependent Variable:** We have identified 'Severity' as the dependent variable as the objective is to understand what factors result in accidents of varying severity.
- **Independent Variables:**
    - From the remaining 46 (forty-six) columns, in our first pass we selected the weather-related columns as the objective was to understand the impact of weather on accident severity. We also added the start time data point by extracting the hour of the day to understand if it impacted the accident severity. Data elements capturing a point of interest ("POI") near the accident site were also retained in the predictor variable set to understand if there was an impact due to their presence nearby. Examples of POI are traffic signal, crossing and junction.

- We determined correlation between the independent weather-related variables.

Temperature and Wind chill had a high correlation, so Wind Chill was dropped off from the predictor variable set.

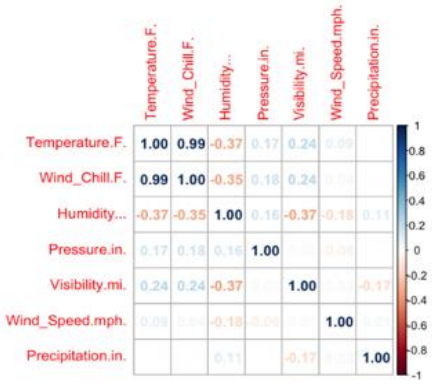

Fig. Correlation Plot of independent weather variables

- POI data elements were captured as binary – True/False. So, we determined the spread between these 2 (two) values across the entire dataset.

Only Traffic Signal, Crossing and Junction data elements had a significant spread between True and False. Our cut-off was to have both the values occur at least 5% of the entire

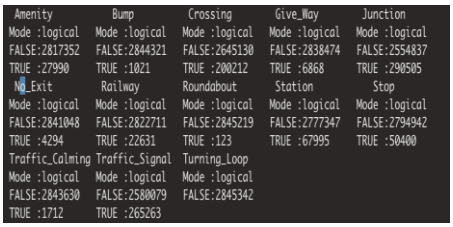

Fig. Summary of POI data elements

- Based on this, our **independent variables** are Start Time (hour of the day), Temperature, Humidity, Visibility, Pressure, Wind Speed, Traffic Signal, Crossing and Junction.

- Next, we performed Exploratory Data Analysis ("EDA") in the independent variables. Some of those analysis points are shared below.

More than 80% of the accidents have a severity of 2 (two).

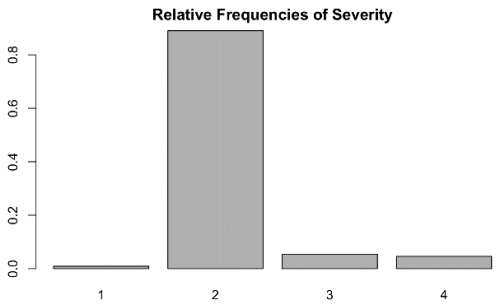

Fig. Relative Frequency of different Severity

Most of the accident sites were contained across the East and the West coasts.

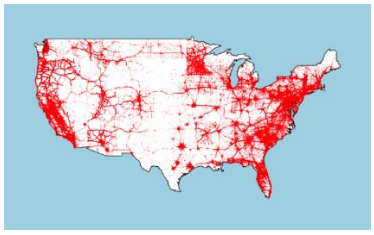

Fig. Relative geographic location of accident sites

Most of the accidents had a low severity.

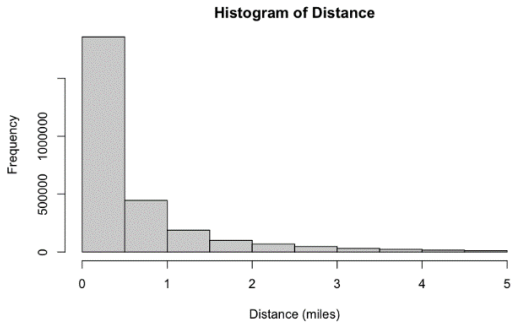

Fig. Length of road impacted due to accidents

Most accident days had moderate temperatures.

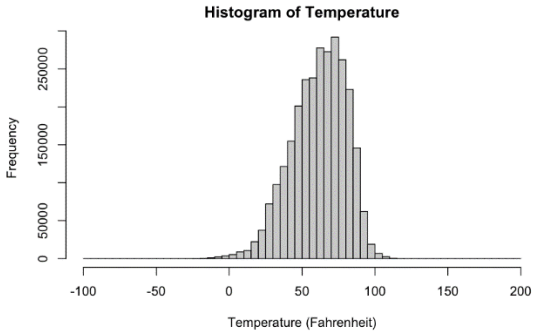

Fig. Temperature Distribution on the day of accidents

Visibility was remarkably high on most of the days when the accidents were recorded.

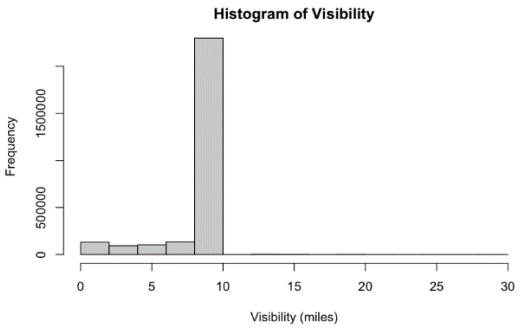

Fig. Visibility Distribution on the day of accidents

- Next, we reviewed the independent variables to identify if they had any missing data or NAs.

NA count was as follows (not captured visually).

- Temperature: 69,274
- Humidity: 73,092
- Pressure: 59,200
- Visibility: 70,546
- Wind Speed: 157,944

We performed data imputation by taking a mean of these data elements and substituting in the place of NAs

Summary Statistics

Variable	N	Mean	Std. Dev.	Min	Pctl. 25	Pctl. 75	Max
Severity	2845342	2.138	0.479	1	2	2	4
Temperature.F.	2776068	61.794	18.623	-89	50	76	196
Humidity...	2772250	64.365	22.875	1	48	83	100
Pressure.in.	2786142	29.472	1.045	0	29.31	30.01	58.9
Visibility.mi.	2774796	9.099	2.718	0	10	10	140
Wind_Speed.mph.	2687398	7.395	5.527	0	3.5	10	1087
Crossing	2845342						
... No	2645130	93%					
... Yes	200212	7%					
Junction	2845342						
... No	2554837	89.8%					
... Yes	290505	10.2%					
Traffic_Signal	2845342						

Fig. Summary Distribution of independent weather variables

- Lastly, zip code in this data set were present both as 5-digit and 9-digit values. So, we standardized all the zip codes as 5-digits to align them with the zip code format in the Income dataset.

Following the above approach, the accidents dataset was ready for analysis.

Income Dataset

- The purpose of this dataset is to use it as a supplemental dataset and append to the Accident dataset. Hence, there is no dependent variable identified within this dataset. There are 19 (nineteen) columns in this dataset. For our analysis, we have only considered 3 (three) columns – Zip Code, Mean Household Income, and the Median Household Income.
- Removed all the records which had a zero mean or a median household income. There were only 303 such records out of a total of more than 15000 records.
- Zip code column was de-duped. Before de-duping, zip codes were grouped by, and the mean of mean household income and the mean of median household income was calculated.
- Next, an EDA was performed on the mean and median household income variables.
- Since this data is from 2011 through 2015, hence the income was adjusted by 14.44% (calculated using CPI for All Urban Consumers (CPI-U) made available by US Bureau of Labour Statistics.) for inflation and extrapolated to 2016-2021 timeframe so that it matches with the Accident dataset.

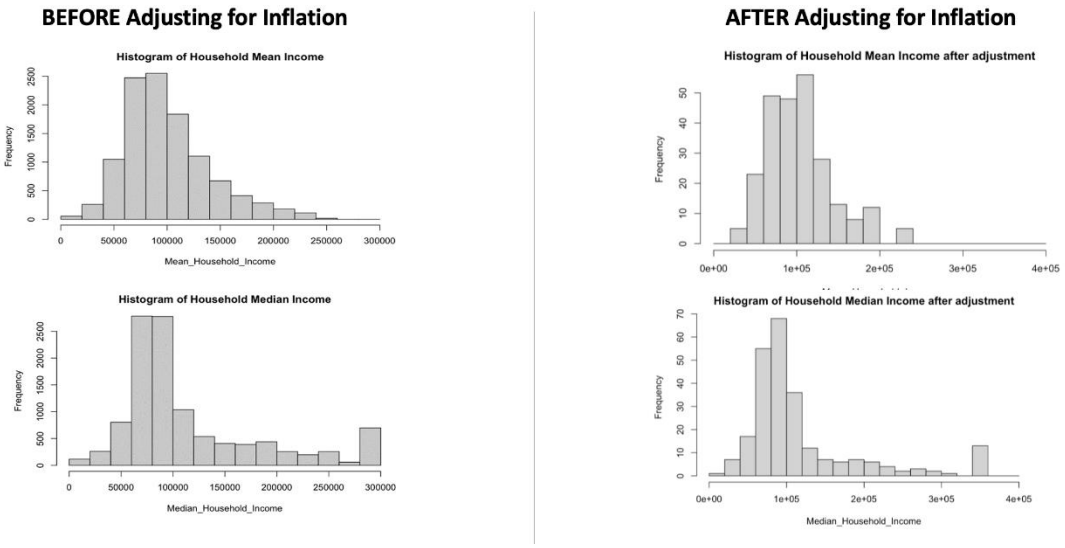

Fig. Income data Histogram

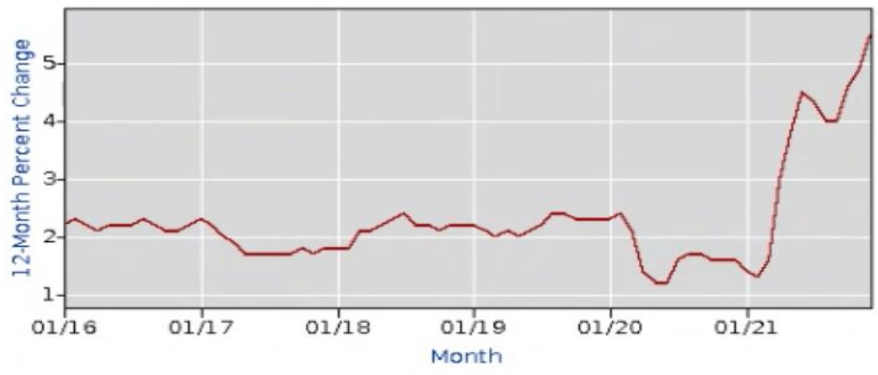

Fig. Inflation data for 2016-2021

Following the above approach, income dataset was ready to be merged into the Accident dataset.

Exploratory Modelling (Initial)

Ran a Random Forest to predict Severity based on environmental and time of day variables (results in Appendix). We sampled only 50,000 records (out of ~3 million) for this first iteration to run the random forest model and check how the processing happens with respect to time it takes and how important the included independent variables are. The sample data frame is partitioned randomly into 70% and 30% training and test data set, respectively.

Random forest consists of many individual decision trees that operate as an ensemble. Each individual tree in the random forest spits out a class prediction and the class with the most votes become our model's prediction. Our model accuracy is 0.8894 - but all predictions are severity "2".

The Kappa Coefficient (-1, 1) states how well the classification performs compared to map. '0' indicates that the classification is random. Less than 0 indicates the classification is significantly worse than random. Greater than 0 (zero) indicates the classification is significantly better than random. Our value of 0.0011 indicates that the classification is random.

This means our model is not a particularly good one in the first iteration we have done. But it mentions which variables are most important in predicting severity which are Pressure, Humidity and Wind Speed. We continued exploring more models with and without changing the independent variables list.

Second model we used is the Gradient Boosting Machine algorithm (GBM). It relies on the intuition that the best possible next model, when combined with previous models, minimizes the overall prediction error. The key idea is to set the target outcomes for this next model to minimize the error. We run the GBM model with the training data previously partitioned and receive the accuracy, kappa, a data frame indicating how important each variable is in determining Severity, and a relative influence table for each variable. It is seen that the accuracy is 89% and the kappa is ~0, so this is not the best model.

Third model is Support Vector Machine (SVM). SVM is a supervised machine learning algorithm that helps in classification and/or regression problems. It aims to find an optimal boundary between the possible outputs. Simply put, SVM does complex data transformations depending on the selected kernel function and based on those transformations, it tries to maximize the separation boundaries between the data points depending on the labels or classes defined. For SVM, only 10,000 records were considered through a random sampling. Only numeric predictor variables were included – Temperature, Visibility, Wind Speed, Humidity and Pressure. We used the "vanilladot" kernel to run our SVM. The model was rejected because the model prediction for Severity was always 2 (two), which represented 89% of the original dataset.

All the initial modelling **results** for Random Forest, GBM and SVM are available in the Appendix.

Feature Engineering

Our initial modelling approach revealed that the Accident dataset had a Class Imbalance problem. Class Imbalance occurs when there is unequal distribution of predicted variable. In our case Severity 2 (class) was occurring in ~89% of the dataset. Hence all our models were skewed towards predicted Severity 2 as the predicted class and hence, irrespective of the complexity and hyperparameter tuning, our models were able to achieve at least 89% accuracy. This seemed correct on the outset, but completely ignores the fact that there are other classes (Severity 1,3 and 4) which are required if we want the model to accurately learn from the dataset.

Hence to solve for Class Imbalance problem, our first approach was to improve the distribution of predicted variable. Another issue that we were facing was the size of the data. At 2.8 million records, our models were taking long time to process and generate results. Hence, we had to sample the data to build and train our models before we could apply the same model to a larger dataset.

We took the following steps to pre-process the data before feeding the various models listed in the previous section:

1. Dropped zip code column from the Accident dataset as it was not a predictor variable for the primary research question.
2. Filtered the dataset to analyse accidents only for Georgia state.
3. To address the imbalanced dataset, we oversampled the minority classes (Severity as 1, 3 or 4) using Synthetic Minority Oversampling Technique ("SMOTE").

SMOTE is used to generate synthetic examples in case of class imbalance problems. Whenever there is a disproportionate difference between the classes, SMOTE tries to generate synthetic records for minority classes (Severity 1,3 and 4 in our case) to bring more balance to the dataset. When this balanced dataset is used for learning the correlation between variables in a dataset, the model does not skew results towards the majority class (Severity 2 in our case).

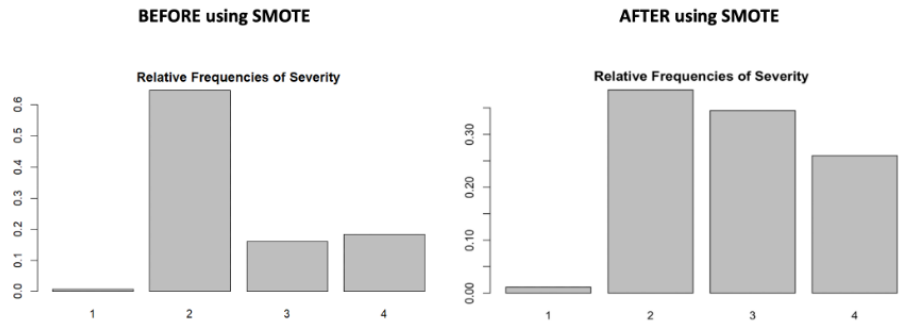

Fig. Severity Distribution comparison before and after applying SMOTE

Final Model Development

The initial modelling techniques used are robust and applicable to the problem statement. Random Forest is primarily used in classification problems due to its ability to combine results from many uncorrelated models. It can manage large data sets containing high number of predictor variables. In our case, we reapplied Random Forecast after performing data manipulation using SMOTE. This time, our Kappa coefficient value of 0.57 indicated that the model was able to identify the relations between the predictor variables much better than the baseline model. Additionally, we were able to identify 'Temperature.F.' and 'Pressure.in.' as variables of importance.

```
Confusion Matrix and Statistics

          Reference
Prediction    1    2    3    4
         1   50    5    3    2
         2  117 6220 1023 1030
         3   64 1043 5351 1205
         4   12  543  741 3108

Overall Statistics

               Accuracy : 0.7179
                 95% CI : (0.7117, 0.724)
    No Information Rate : 0.3807
    P-Value [Acc > NIR] : < 2.2e-16

                  Kappa : 0.5718

 Mcnemar's Test P-Value : < 2.2e-16

Statistics by Class:

                     Class: 1 Class: 2 Class: 3 Class: 4
Sensitivity          0.205761   0.7963   0.7518   0.5815
Specificity          0.999507   0.8292   0.8274   0.9146
Pos Pred Value       0.833333   0.7414   0.6983   0.7057
Neg Pred Value       0.990566   0.8688   0.8625   0.8612
Prevalence           0.011844   0.3807   0.3469   0.2605
Detection Rate       0.002437   0.3032   0.2608   0.1515
Detection Prevalence 0.002924   0.4089   0.3735   0.2147
Balanced Accuracy    0.602634   0.8128   0.7896   0.7480
```

	Overall <dbl>
Temperature.F.	6792.300
Humidity...	5884.261
Pressure.in.	8991.274
Visibility.mi.	1541.228
Wind_Speed.mph.	5360.761

Fig. Output of the Random Forest Model

The second model that we reapplied for Gradient Boosting Machine ("GBM"). GBMs are primarily used for classification problems as this ensemble technique involves building a stronger model using a collection of weaker models. The loss function in GBM optimizes the overall model by minimizing the gap between actual and predicted values. Each iteration in the GBM should reduce the value of our loss function. After rebalancing the dataset using SMOTE, we applied GBM and were able to achieve a Kappa coefficient value of 0.338. It also helped us identify 'Pressure.in' and 'Wind_Speed.mph' as variables of importance.

```
Confusion Matrix and Statistics

          Reference
Prediction    1    2    3    4
         1    2    2    0    0
         2  154 5493 1341 1491
         3   82 1714 5002 2657
         4    5  602  775 1197

Overall Statistics

               Accuracy : 0.57
                 95% CI : (0.5632, 0.5768)
    No Information Rate : 0.3807
    P-Value [Acc > NIR] : < 2.2e-16

                  Kappa : 0.3384

 Mcnemar's Test P-Value : < 2.2e-16

Statistics by Class:

                     Class: 1 Class: 2 Class: 3 Class: 4
Sensitivity          8.230e-03   0.7032   0.7027  0.22395
Specificity          9.999e-01   0.7650   0.6677  0.90891
Pos Pred Value       5.000e-01   0.6478   0.5290  0.46413
Neg Pred Value       9.883e-01   0.8074   0.8087  0.76876
Prevalence           1.184e-02   0.3807   0.3469  0.26052
Detection Rate       9.748e-05   0.2677   0.2438  0.05834
Detection Prevalence 1.950e-04   0.4133   0.4608  0.12570
Balanced Accuracy    5.041e-01   0.7341   0.6852  0.56643
```

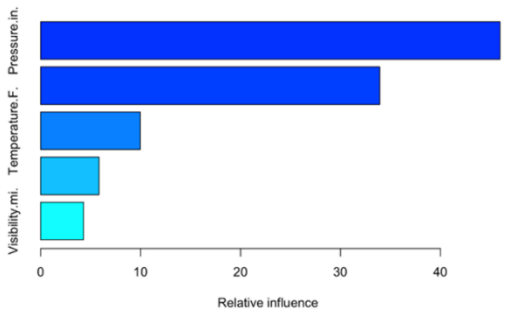

Fig. Output of the GBM Model

Finally, we applied the Support Vector Machine (SVM) model after rebalancing the data using SMOTE. SVM can manage both linear and non-linear data. Hence it is still applicable for our problem statement. When we reapplied the SVM model to SMOTE processed data, we were able to overcome the problem of default prediction. In the first iteration, the majority class (Severity 2) was being predicted for all test data which resulted in ~89% accuracy. However, this was incorrect since the Severity for almost 90% of data as 2 (two). In second iteration, we were able to achieve other Severity classes and the final accuracy of SVM model was 50.8%.

```
> svm_model <- ksvm(Severity ~ Temperature.F. + Humidity... + Pressure.in. + Visibility.mi. +Wind_Speed.mph., d
ata = training, type = 'C-svc', kernel = 'vanilladot')
 Setting default kernel parameters
> # svm_model
> a <- colSums((svm_model@xmatrix[[1]]*svm_model@coef[[1]]))
> a0 <- -svm_model@b
> a
 Temperature.F.     Humidity...    Pressure.in.  Visibility.mi. Wind_Speed.mph.
   1.006133e-05   -1.415759e-05   -6.758297e-06   -5.582721e-05   -1.059377e-05
> a0
[1]  0.99997342  0.99994699  0.99994822  0.04566801 -0.06623489 -0.99984911
> pred <- predict(svm_model, testing)
> # pred
> distinct_pred <- unique(pred)
> distinct_pred
[1] 3 2
Levels: 1 2 3 4
> sum(pred == testing$Severity)/nrow(testing)
[1] 0.5089438
```

Fig. Output of the SVM Model

Secondary Research Questions - Analysis

- **Question 1: How is the severity of US road accident linked to the socio-economic factor i.e., median income of the population in a zip code.**
    a. We analyzed this question using EDA technique.
    b. First, we adjusted the mean and median income with the inflation data as described in the Data Wrangling section for the Income dataset.
    c. Subsequently, we plotted a graph showing severity against mean and median income.

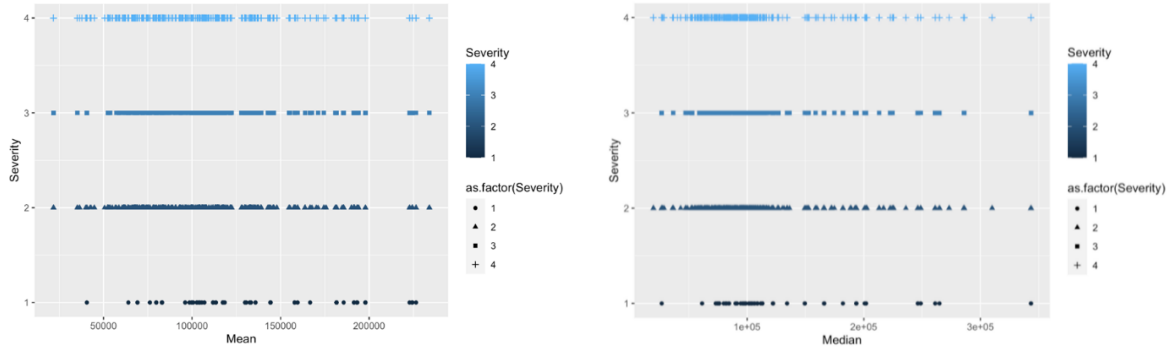

Fig. Distribution of Accident Severity across Mean and Median neighbourhood incomes

- **Question 2: How did the severity of US road accident change between the pre-Covid and the Covid period.**
    a. We analyzed this question using EDA technique.
    b. We plotted a graph showing the frequency of different severity across the 2 (two) periods.

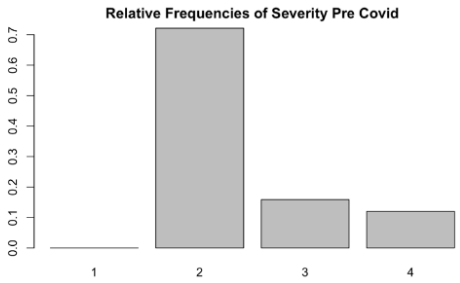

Fig. Accident Severity distribution during pre-Covid period

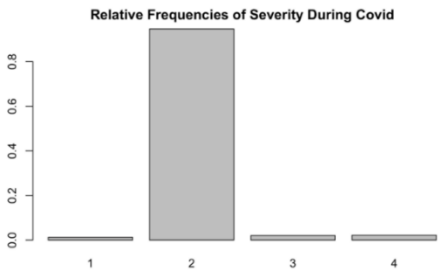

Fig. Accident Severity distribution during Covid period

Business Impact

We are linking this to the Business Justification section discussed earlier in this report. Auto-insurance companies can use the findings of the report to incentivize and educate their customers about the hazardous driving conditions linked to the weather parameters. Companies can use the weather data and the real-time location of where the customer is driving (assuming they have this data) to send real-time push notifications of any hazardous driving conditions that the model predicts.

Conclusion

- **Primary Research Question - How is the US road accident severity linked to the prevalent weather at the time of the accident?**
  Based on the 3 (three) modelling algorithms that were executed - Random Forest, GBM and SVM, output from Random Forest was the best with a kappa of 0.57 and an accuracy of 71.8%. Top two (2) predictor variables observed from Random Forest were Pressure and Temperature. This finding relates to our initial hypothesis that the weather-related parameters have a correlation on the severity of an accident.
- **Secondary Research Question 1 - How is the US road accident severity linked to the socio-economic factor i.e., the median income of the population residing in an area, defined by a zip code?**
  There was no correlation between the Severity and the mean and the median income of the neighbourhood. No correlation between Severity and mean/median income of the neighbourhood finding did not align with our initial hypothesis which was that the low-income neighbourhoods may witness more severe accidents.
- **Secondary Research Question 2 - How is the US road accident severity linked to the Covid-related lockdowns?**
  The number of instances low severity accidents (1 or 2) increased during Covid compared to the earlier period. This finding did not align with our initial hypothesis which was that the accidents severity did not change between both the time periods. Further, we think that this may be due to lesser traffic congestion on the roads during Covid which would result in a lesser stretch of road getting affected in the event of an accident.

The analysis was limited to Georgia state only due to the volume constraint of the dataset; same modelling technique can be used for all the other records within the Accident dataset.

Appendix

Random Forest – Initial Modelling result:

```
Confusion Matrix and Statistics

          Reference
Prediction     1     2     3     4
         1     1     0     0     0
         2   157 13338   828   674
         3     0     0     0     0
         4     0     0     0     0

Overall Statistics

               Accuracy : 0.8894
                 95% CI : (0.8843, 0.8944)
    No Information Rate : 0.8893
    P-Value [Acc > NIR] : 0.4962

                  Kappa : 0.0011

 Mcnemar's Test P-Value : NA

Statistics by Class:

                     Class: 1  Class: 2 Class: 3 Class: 4
Sensitivity          6.329e-03 1.0000000  0.00000  0.00000
Specificity          1.000e+00 0.0006024  1.00000  1.00000
Pos Pred Value       1.000e+00 0.8893779      NaN      NaN
Neg Pred Value       9.895e-01 1.0000000  0.94479  0.95506
Prevalence           1.053e-02 0.8893186  0.05521  0.04494
Detection Rate       6.668e-05 0.8893186  0.00000  0.00000
Detection Prevalence 6.668e-05 0.9999333  0.00000  0.00000
Balanced Accuracy    5.032e-01 0.5003012  0.50000  0.50000
```

	Overall <dbl>
Start_Time01	2.756089
Start_Time02	2.456127
Start_Time03	3.180122
Start_Time04	3.801493
Start_Time05	3.019628
Start_Time06	2.789163
Start_Time07	5.096381
Start_Time08	3.659653
Start_Time09	3.394632
Start_Time10	2.778206
Start_Time11	3.442350
Start_Time12	3.383920
Start_Time13	3.811090
Start_Time14	3.756088
Start_Time15	3.474836
Start_Time16	3.521394
Start_Time17	3.689862
Start_Time18	4.014008
Start_Time19	3.353767
Start_Time20	3.912329
Start_Time21	2.968733
Start_Time22	3.455419
Start_Time23	2.970202
Temperature.F.	47.630883
Humidity...	46.836035
Pressure.in.	81.590710
Visibility.mi.	27.010862
Wind_Speed.mph.	75.322441
CrossingTRUE	8.618846
JunctionTRUE	17.799846
Traffic_SignalTRUE	15.285157

Gradient Boosting Machine - Initial Modelling result:

```
Confusion Matrix and Statistics

          Reference
Prediction     1     2     3     4
         1     0     0     0     0
         2   140 13371   815   670
         3     0     0     2     0
         4     0     0     0     0

Overall Statistics

               Accuracy : 0.8917
                 95% CI : (0.8866, 0.8966)
    No Information Rate : 0.8915
    P-Value [Acc > NIR] : 0.4857

                  Kappa : 0.0023

 Mcnemar's Test P-Value : NA

Statistics by Class:

                     Class: 1 Class: 2  Class: 3 Class: 4
Sensitivity          0.000000 1.000000 0.0024480  0.00000
Specificity          1.000000 0.001229 1.0000000  1.00000
Pos Pred Value            NaN 0.891638 1.0000000      NaN
Neg Pred Value       0.990665 1.000000 0.9456522  0.95533
Prevalence           0.009335 0.891519 0.0544739  0.04467
Detection Rate       0.000000 0.891519 0.0001334  0.00000
Detection Prevalence 0.000000 0.999867 0.0001334  0.00000
Balanced Accuracy    0.500000 0.500615 0.5012240  0.50000
```

	var <chr>	rel.inf <dbl>
Pressure.in.	Pressure.in.	45.760904
Wind_Speed.mph.	Wind_Speed.mph.	33.192491
Humidity...	Humidity...	7.545162
Visibility.mi.	Visibility.mi.	6.920665
Temperature.F.	Temperature.F.	6.580777

Support Vector Machine – Initial Modelling result:

```
> svm_model <- ksvm(Severity ~ ., data = training, type = 'C-svc', kernel = 'vanilladot')
 Setting default kernel parameters
> a <- colSums((svm_model@xmatrix[[1]]*svm_model@coef[[1]]))
> a0 <- -svm_model@b
> a
 Temperature.F.     Humidity...    Pressure.in.  Visibility.mi. Wind_Speed.mph.
  -5.032781e-05   -3.973127e-05    1.194562e-05   -1.572706e-06   -4.643679e-06
> a0
[1]  0.9999922  1.0000304  0.9999764 -1.0000155 -1.0000163 -0.9999114
> pred <- predict(svm_model, testing)
> distinct_pred <- unique(pred)
> distinct_pred
[1] 2
Levels: 1 2 3 4
```

References and Data Sources

References

- Accident Risk Prediction based on Heterogeneous Sparse Data: New Dataset and Insights

  https://arxiv.org/pdf/1909.09638.pdf

- A Countrywide Traffic Accident Dataset

  https://arxiv.org/pdf/1906.05409.pdf

- Research Paper suggesting using SMOTE to address imbalanced dataset

  https://arxiv.org/pdf/1106.1813.pdf

Data Sources

- US Accidents (2016 - 2021)
  https://www.kaggle.com/datasets/sobhanmoosavi/us-accidents
- US Household Income Statistics
  https://www.kaggle.com/datasets/goldenoakresearch/us-household-income-stats-geo-locations/versions/1
- US Inflation Data
  https://data.bls.gov/