Severe Car Accidents Prediction

Kai-Jo Ma

Department of Computer Science, Virginia Tech

CS 5525: Data Analytics

Dr. Reza Jafari

December 11, 2022

Table of Contents

Table of Figures

Table of Tables

Abstract

The main goal of this project is to learn the patterns of a severe car accident in the United States and predict whether a car accident is severe when it occurs. The dataset is collected through various transport departments from 2016 through 2021.

We apply 7 different machine learning classifiers, which are Logistic Regression, SVM, KNN, Decision Tree, Naïve Bayes, Random Forest, and Neural Network respectively.

Random Forest has the best performance on our dataset, with accuracy of 0.873, precision of 0.858, recall of 0.883, specificity of 0.864, and F1 score of 0.870. In addition, Random Forest performs best on ROC curve and AUC as well, with AUC of 0.941. Therefore, we recommend Random Forest on our car accident dataset.

In conclusion, reducing car accidents is important since car accidents have large economical and societal impacts on us, especially the severe ones. By analyzing and acquiring the patterns of severe car accidents, we could be able to decrease tragedy from happening on the roads.

Introduction

This project is aimed to implement machine learning classifiers on real word dataset. The dataset we used is regarding car accidents in the United States. Our goal is to predict whether a car accident is severe or not when it occurs.

There are mainly three phases in this project. The first phase is data preprocessing. First, we define our target and attributes. Then we will do Exploratory Data Analysis (EDA) on our dataset. We apply some data visualizations so that we can understand our dataset better and make complicated data simpler. We observe the distribution of our target and handle the imbalance problem in our dataset. After that, we will clean our dataset and deal with missing values reasonably. Next, we will select features that will be trained in machining learning classifiers in the third phase. Lastly, we will make necessary transformations on our dataset such as binary encoding and standardization.

The second phase is regression analysis. We utilize backward stepwise regression on our dataset and observe its performance. In addition, we also apply F test and confidence interval analysis to observe the variances of our variables. We also conduct collinearity analysis to check the multicollinearity between our variables.

The third phase is machine learning analysis. We will use the dataset which has been pre-processed in the previous phases on various machine learning classifiers. We will visualize the performances of each machine learning classifier, compare them through different metrics, and finally summarize our work and recommend a machine learning classifier on our dataset.

Description of the dataset

This dataset about car accidents in the United States includes data from all 49 states. It contains roughly 2.8 million accident records (2,845,342 observations) and 47 features, where 14 of them are numerical and 33 of them are categorical. The time period where car accident data is collected is from February 2016 to December 2021 and it is not a time series data. The source of accident data comes from the US and state departments of transportation, law enforcement organizations, traffic cameras, and traffic sensors embedded on the road.

Severity serves as the dependent variable while the other attributes serve as independent variables in the car accident dataset. The car accident dataset is important because cars are everywhere in our lives either as a car owner or a commuter. Having a car accident often causes huge economical and societal costs. Among all the costs car accidents have brought about, a large portion of the cost is caused by a relatively small number of serious accidents. Therefore, decreasing car accidents is undoubtedly significant, especially decreasing the severe ones. One proactive way to reduce car accidents is to prevent them from happening. Therefore, the goal of this project is to predict the severity of car accidents. If we can know the patterns and factors of severe car accidents, we may be able to be cautious of them and avoid them.

Phase I

Since the objective of this project is to predict the severity of a car accident, we will first analyze the severity feature. The severity feature is our target and the other features are our attributes.

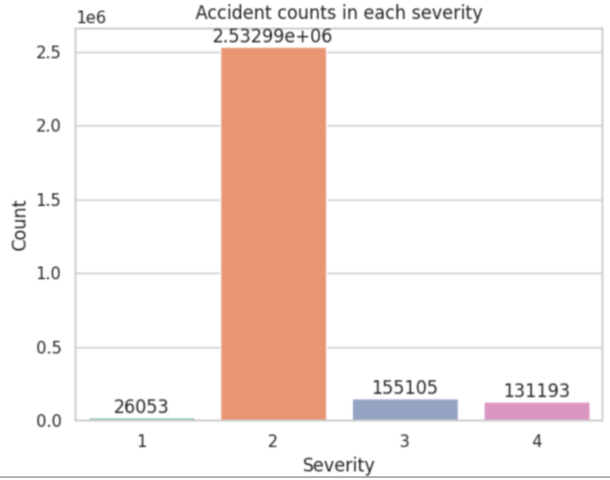


Figure 1: Accident counts in each severity level

There are four levels of severity, from one to four, where one is the least severe and four indicates the most severe car accident. From the above figure, we can observe that the dataset is very imbalanced. The severity level two dominates in the dataset and the count is much larger than other levels. Therefore, we will solve the imbalance problem before we train our machining learning classifiers.

Next, we are interested in the duration of car accidents with each severity level. The average duration of a car accident in this dataset is around 359 minutes, which is around 6 hours.

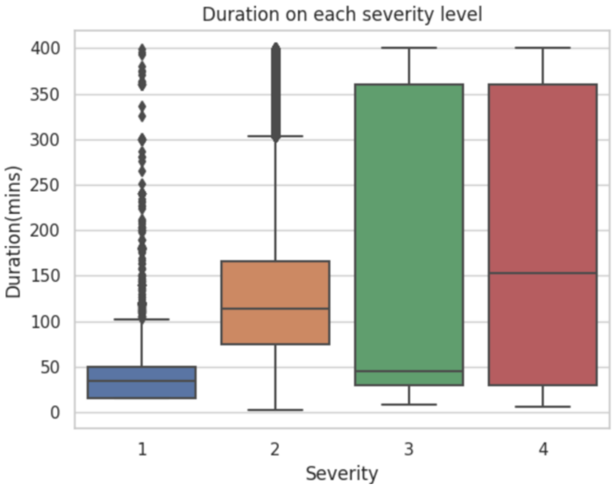


Figure 2: Duration of each severity level

From the box plot of duration versus severity, we can see that car accidents with severity level 1 last within an hour, which is quite reasonable since they are just fender benders. Interestingly, the durations of car accidents with the largest severity level ranges widely. Therefore, we can conclude that serious car accidents can occur in a really short time and it is really important that we analyze the key factors which result in serious car accidents.

Since the dataset is collected from all the states in the United States, we could analyze the accident counts in each state.

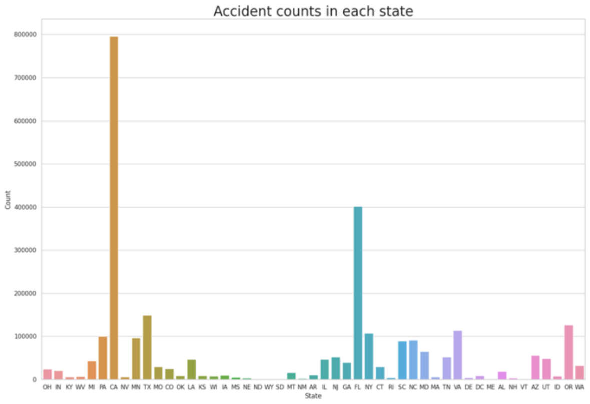


Figure 3: Accident counts in each state

The top 5 states which have the most car accidents are California, Florida, Texas, Oregon, and Virginia, with 795868, 401,388, 149,037, 126,341, and 113,535 car accidents respectively.

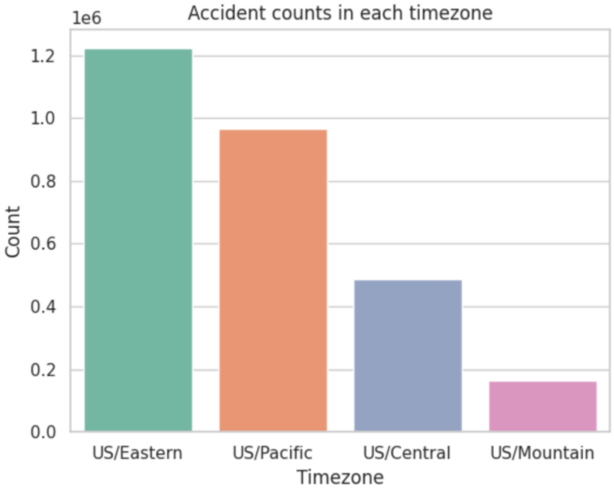
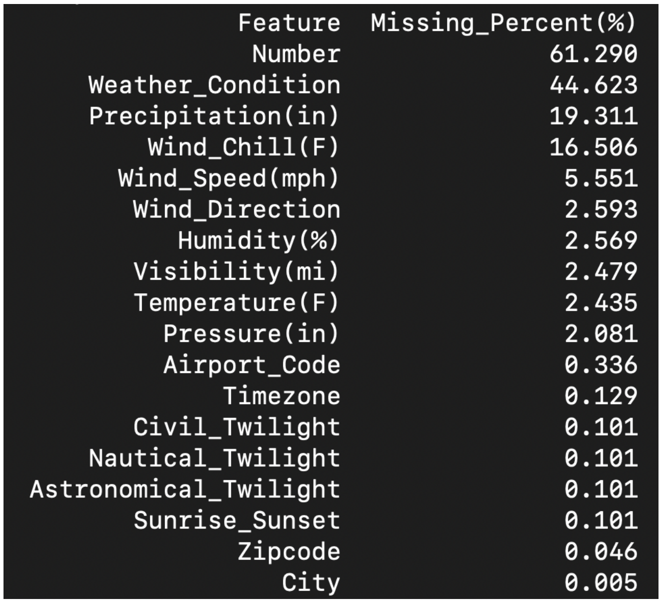


Figure 4: Accident counts in each time zone

As for car accidents in each time zone, the Eastern time zone has the most car accidents and the Mountain time zone has the least. This observation is aligned with the population in United States. Within the United States, the Eastern time zone is the most populous region while the Mountain time zone is the least populous region.

Now, we select features which would be applied in machine learning classifiers and eliminate unnecessary ones. The feature selection process contains three parts. First, eliminate features that are useless, and second, deal with missing values and the last step is to reduce dimensions with low feature importance by Random Forest.
First, the IDs and descriptions of car accidents are not relevant to the severity level, so they can be removed. Distance which car accidents cause can only be collected after the car accident occurs. Since our objective is to predict the severity level of a car accident, we won't know this distance result in advance, and therefore it should be removed. Similarly, end time, end latitude, end longitude features can only be acquired after the car accident and should be removed as well. Next, we want to drop categorical features which have only one unique value since having only one class isn't helpful to the learning of machine learning classifiers. Therefore, we remove the country feature and turning loop feature where the turning loop feature indicates the presence of turning loops and it is always false in our dataset.

The second step is to deal with missing values.



```
                 Feature  Missing_Percent(%)
                  Number              61.290
       Weather_Condition              44.623
       Precipitation(in)              19.311
           Wind_Chill(F)              16.506
         Wind_Speed(mph)               5.551
          Wind_Direction               2.593
             Humidity(%)               2.569
          Visibility(mi)               2.479
          Temperature(F)               2.435
            Pressure(in)               2.081
            Airport_Code               0.336
                Timezone               0.129
          Civil_Twilight               0.101
       Nautical_Twilight               0.101
   Astronomical_Twilight               0.101
          Sunrise_Sunset               0.101
                 Zipcode               0.046
                    City               0.005
```

Figure 5: Missing value ratio of features

From the missing values and missing ratios analysis in each column, we can observe that the missing ratios of number, weather condition, precipitation, and wind chill features are rather high. Number indicates the street numbers in the address fields. We first drop number and wind chill features. For precipitation, since we believe the amount of precipitation is critical in predicting the severity level of a car accident, we fill the empty values in precipitation with the median of it. As for weather condition, in addition to the high missing ratios, there are too many categories in it while some of them refer to similar weather conditions. For example, heavy rain, rain shower, heavy t-storm, heavy thunderstorms can be classified in the heavy storm category, and snow, sleet, ice can be classified as snow. Therefore, we put similar weather conditions in the same class to reduce the number of categories, reducing the number of categories in weather condition from 18 to 7. Finally, we drop the weather condition feature and add 7 new features, clear, cloud, rain, heavy rain, snow, heavy snow, and fog and analyze the car accident counts in different weather conditions.

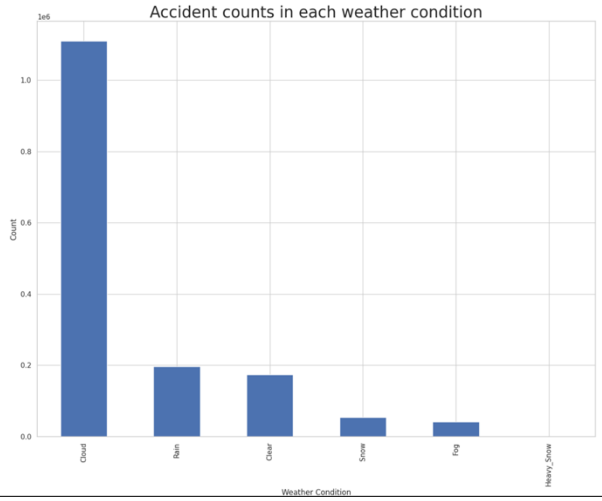


Figure 6: Accident counts in each weather condition

The last method of dimension reduction is to use random forest to calculate the feature importance. We split our data into 80% and 20% portions, where training data accounts for 80% and testing data accounts for 20% and the split data would be applied in all the machining learning classifiers.

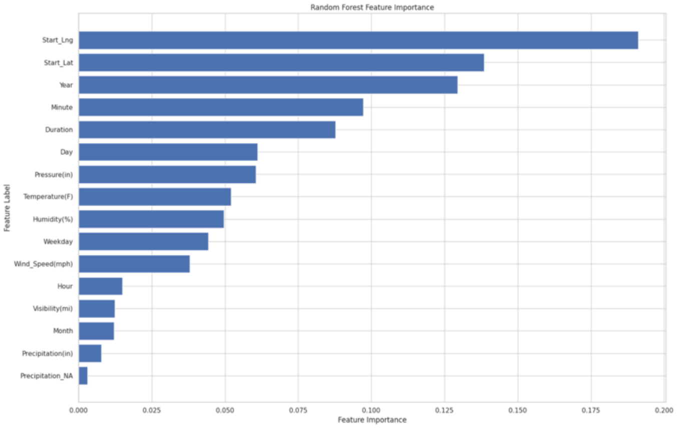


Figure 7: Random Forest feature importance

We drop the precipitation_NA feature since its feature importance is quite little.
It is worth noting that there are too many categories in the wind feature, similar to the original weather condition feature. Therefore, we gather similar wind directions into the same category. For example, South, SSW, SSE are renamed as S, and so on. Below is the analysis of car accident counts in different wind conditions.

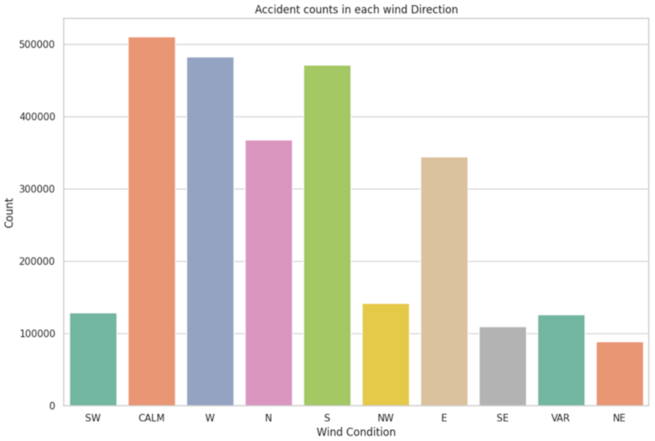


Figure 8: Accident counts in each wind direction

In the original dataset, there is a start time feature which is in the format of “YYYY-MM-DD H:M:S”. We split this format and create new features of year, month, day, weekday, hour, and minute and delete the original start time feature in order to make models learn better.
The sample covariance matrix created through heat map is as follows.

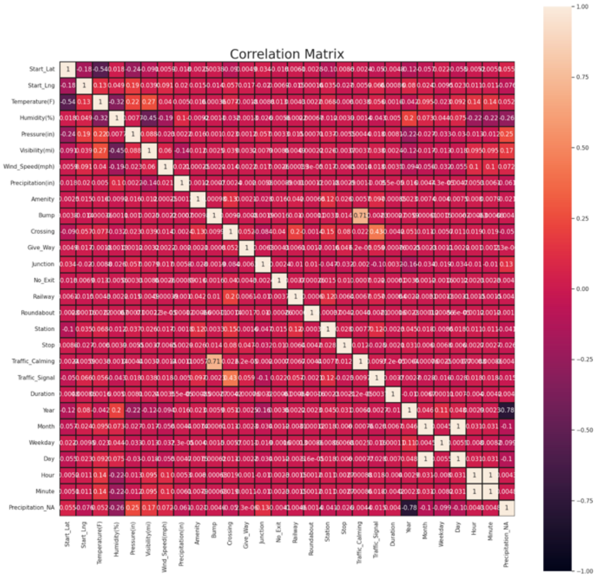
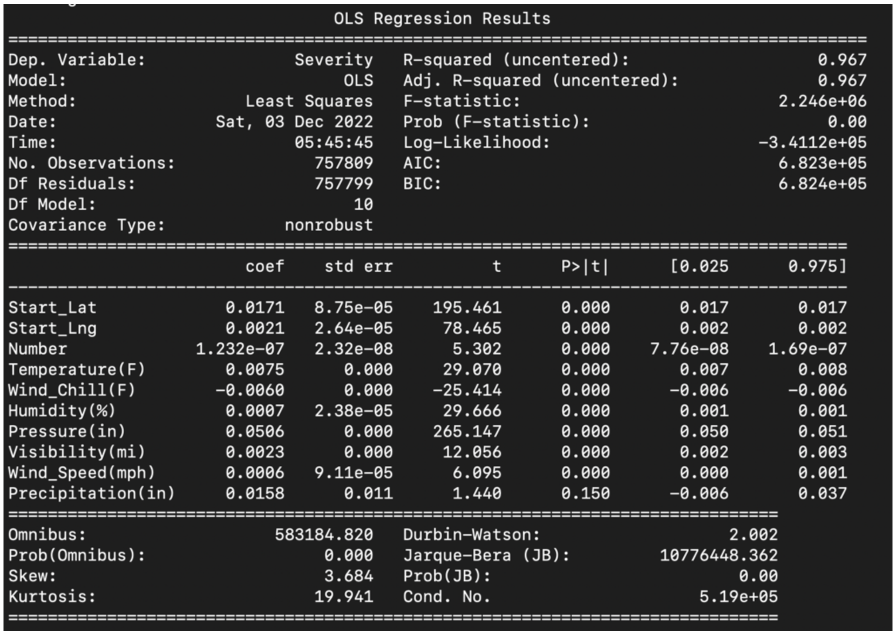


Figure 9: Correlation matrix of features

Lastly, as mentioned previously, our dataset is imbalanced, with severity level 2 dominating in the severity levels. Therefore, we under sample level 2 to 10% of its original number in order to balance the severity levels. We also standardize our dataset and in terms of feature encoding, we transform categorical features into numerical ones by using binary encoding.

Phase II

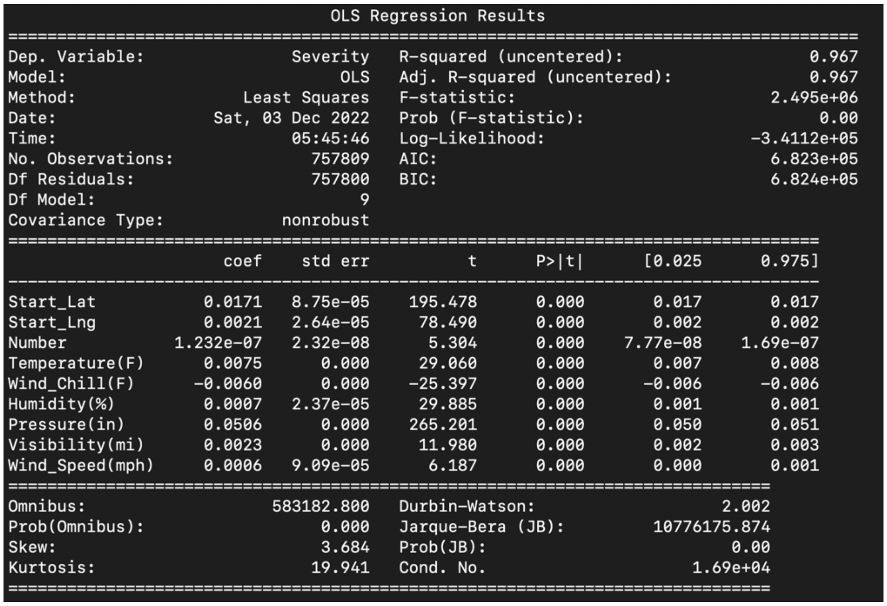
In this phase, we will apply regression analysis. The below is the summary of the ordinary least square regression on our dataset.



```
                                 OLS Regression Results
=======================================================================================
Dep. Variable:               Severity   R-squared (uncentered):                   0.967
Model:                            OLS   Adj. R-squared (uncentered):              0.967
Method:                 Least Squares   F-statistic:                          2.246e+06
Date:                Sat, 03 Dec 2022   Prob (F-statistic):                        0.00
Time:                        05:45:45   Log-Likelihood:                     -3.4112e+05
No. Observations:              757809   AIC:                                  6.823e+05
Df Residuals:                  757799   BIC:                                  6.824e+05
Df Model:                          10
Covariance Type:            nonrobust
=====================================================================================
                        coef    std err          t      P>|t|      [0.025      0.975]
-------------------------------------------------------------------------------------
Start_Lat             0.0171   8.75e-05    195.461      0.000       0.017       0.017
Start_Lng             0.0021   2.64e-05     78.465      0.000       0.002       0.002
Number             1.232e-07   2.32e-08      5.302      0.000    7.76e-08    1.69e-07
Temperature(F)        0.0075      0.000     29.070      0.000       0.007       0.008
Wind_Chill(F)        -0.0060      0.000    -25.414      0.000      -0.006      -0.006
Humidity(%)           0.0007   2.38e-05     29.666      0.000       0.001       0.001
Pressure(in)          0.0506      0.000    265.147      0.000       0.050       0.051
Visibility(mi)        0.0023      0.000     12.056      0.000       0.002       0.003
Wind_Speed(mph)       0.0006   9.11e-05      6.095      0.000       0.000       0.001
Precipitation(in)     0.0158      0.011      1.440      0.150      -0.006       0.037
==============================================================================
Omnibus:                   583184.820   Durbin-Watson:                   2.002
Prob(Omnibus):                  0.000   Jarque-Bera (JB):         10776448.362
Skew:                           3.684   Prob(JB):                         0.00
Kurtosis:                      19.941   Cond. No.                     5.19e+05
==============================================================================
```

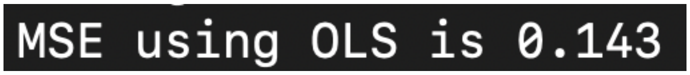
Figure 10: Initial OLS regression results

The original adjusted R squared is 0.967. We use backward stepwise regression and remove one predictor at a time. We first remove the precipitation feature since the p-value of it on t test is the largest.



```
                                 OLS Regression Results
=======================================================================================
Dep. Variable:               Severity   R-squared (uncentered):                   0.967
Model:                            OLS   Adj. R-squared (uncentered):              0.967
Method:                 Least Squares   F-statistic:                          2.495e+06
Date:                Sat, 03 Dec 2022   Prob (F-statistic):                        0.00
Time:                        05:45:46   Log-Likelihood:                     -3.4112e+05
No. Observations:              757809   AIC:                                  6.823e+05
Df Residuals:                  757800   BIC:                                  6.824e+05
Df Model:                           9
Covariance Type:            nonrobust
===================================================================================
                      coef    std err          t      P>|t|      [0.025      0.975]
-----------------------------------------------------------------------------------
Start_Lat           0.0171   8.75e-05    195.478      0.000       0.017       0.017
Start_Lng           0.0021   2.64e-05     78.490      0.000       0.002       0.002
Number           1.232e-07   2.32e-08      5.304      0.000    7.77e-08    1.69e-07
Temperature(F)      0.0075      0.000     29.060      0.000       0.007       0.008
Wind_Chill(F)      -0.0060      0.000    -25.397      0.000      -0.006      -0.006
Humidity(%)         0.0007   2.37e-05     29.885      0.000       0.001       0.001
Pressure(in)        0.0506      0.000    265.201      0.000       0.050       0.051
Visibility(mi)      0.0023      0.000     11.980      0.000       0.002       0.003
Wind_Speed(mph)     0.0006   9.09e-05      6.187      0.000       0.000       0.001
==============================================================================
Omnibus:                   583182.800   Durbin-Watson:                   2.002
Prob(Omnibus):                  0.000   Jarque-Bera (JB):         10776175.874
Skew:                           3.684   Prob(JB):                         0.00
Kurtosis:                      19.941   Cond. No.                     1.69e+04
==============================================================================
```

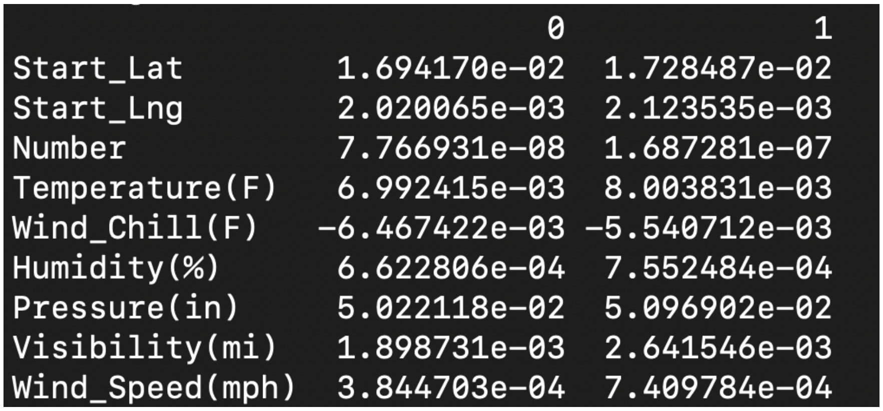
Figure 11: OLS regression results after backward stepwise regression



```
MSE using OLS is 0.143
```

Figure 12: Mean square error of final model

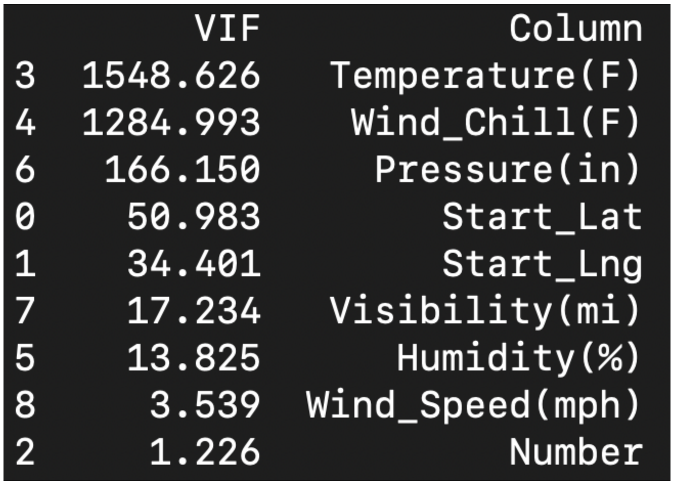
After removing the precipitation feature, the p-values by t test of all the other features are zero, so we stop the backward stepwise regression. The adjusted R squared is 0.967 and the mean square error between true value and predicted value is 0.143. In addition, the p-values of the F test are smaller than the threshold, so we can reject the null hypothesis that the fit of the intercept-only model and our model is equal. In other words, our model provides a better fit than the intercept-only model.
Next, we analyze the confidence intervals with 95% confidence on our final predictors.



```
                              0             1
Start_Lat         1.694170e-02  1.728487e-02
Start_Lng         2.020065e-03  2.123535e-03
Number            7.766931e-08  1.687281e-07
Temperature(F)    6.992415e-03  8.003831e-03
Wind_Chill(F)    -6.467422e-03 -5.540712e-03
Humidity(%)       6.622806e-04  7.552484e-04
Pressure(in)      5.022118e-02  5.096902e-02
Visibility(mi)    1.898731e-03  2.641546e-03
Wind_Speed(mph)   3.844703e-04  7.409784e-04
```

Figure 13: Confidence analysis of features

Lastly, we apply collinearity analysis and detect multicollinearity in the independent variables on our dataset. The collinearity analysis shows that most of the predictors in our dataset are highly correlated.



```
        VIF           Column
3  1548.626   Temperature(F)
4  1284.993    Wind_Chill(F)
6   166.150     Pressure(in)
0    50.983        Start_Lat
1    34.401        Start_Lng
7    17.234   Visibility(mi)
5    13.825      Humidity(%)
8     3.539  Wind_Speed(mph)
2     1.226           Number
```

Figure 14: Collinearity analysis of features

Phase III

In this phase, we will apply various machine learning classifiers on the car accident datasets to predict whether a car accident is severe. Since our main goal is to acquire learning patterns and causes of severe car accidents, we transform our task into a binary classification problem. Specifically, we would like to know whether a car accident is severe when it occurs; therefore, severity levels 1 and 2 are categorized as not severe and severity level 3 and 4 are categorized as severe.

We apply 7 different machine learning classifiers on our dataset, logistic regression, support vector machine, naïve bayes, k-nearest neighbor, decision tree, random forest, and neural network. The performances of each machine learning classifier are listed below. Note that while training SVM, it can hardly converge due to our large training data size.

Table 1: Performances of Logistic Regression, SVM, and KNN

	Logistic Regression	**SVM**	**KNN**
Accuracy	0.801	0.799	0.800
Precision	0.803	0.807	0.790
Recall	0.777	0.768	0.798
Specificity	0.823	0.829	0.803
F1 score	0.790	0.787	0.794
Best Parameters	{'C': 10}	{'C': 1}	{'n_neighbors': 10, 'p': 1, 'weights': 'distance'}

Table 2: Performances of Decision Tree, Naive Bayes, Random Forest, and Neural Network

	Decision Tree	**Naïve Bayes**	**Random Forest**	**Neural Network**
Accuracy	0.849	0.728	0.873	0.853
Precision	0.836	0.699	0.858	0.842
Recall	0.855	0.767	0.883	0.854
Specificity	0.844	0.693	0.864	0.851
F1 score	0.846	0.731	0.870	0.848
Best Parameters	{'min_samples_leaf': 0.0002, 'max_features': None, 'max_depth': 80, 'criterion': 'entropy'}		{'n_estimators': 200, 'min_samples_leaf': 5, 'max_features': None, 'max_depth': 80, 'criterion': 'entropy'}	{'solver': 'adam', 'learning_rate_init': 0.01, 'hidden_layer_sizes': (128, 64), 'batch_size': 512, 'alpha': 0.01}

We apply confusion matrices to show how different classes are classified and misclassified by these machine learning classifiers.

We also apply ROC curve and AUC to compare the performances of each machine learning classifier.

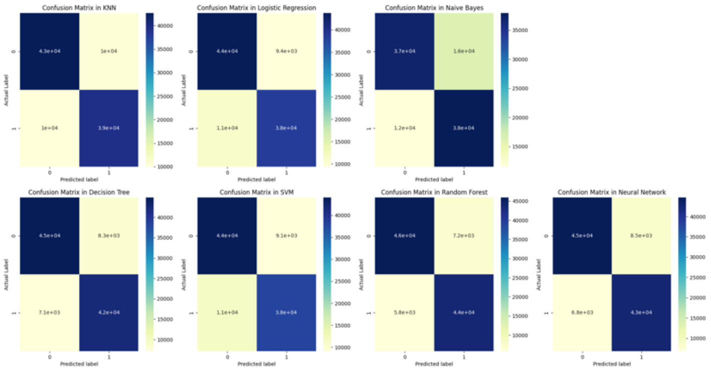


Figure 15: Confusion matrix of each machine learning classifiers

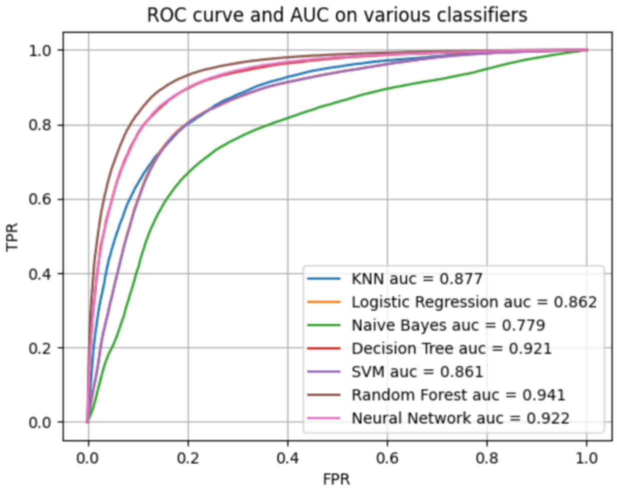


Figure 16: ROC curve and AUC on each machine learning classifier

From the ROC curve and AUC, we can observe that Random Forest has the best performance. Random forest performs well since it reduces overfitting problems by creating multiple trees. Having multiple trees can reduce variance since the trees in random forest are averaged. As a result, random forest performs accurate and precise results and reduces overfitting. Naïve Bayes has the worst performance since it assumes that all the attributes are mutually independent. However, in the real dataset, this is almost impossible. In addition, in terms of single model, neural network and decision tree are relatively better than the others.

Recommendation

Random Forest performs best on our car accident dataset with the most upper left ROC curve and the highest AUC value.

Contribution and Conclusion

We apply EDA, feature importance, and so on to understand and process our features. We also apply backward stepwise regression to select independent variables with respect to car accident severity. Last but not least, we compare the performance by various evaluation metrics and highlight the superiority of Random Forest.
In our dataset, Random Forest performs best Naïve Bayes performs worst. Therefore, we recommend Random Forest on our dataset and our goal, predict whether it is severe when a car accident occurs. With this predictor on sever car accidents, we could be able to decrease tragedy from happening on the roads.

Future Work

- More powerful tree-based models can be considered, such as XGBoost or LightGBM.
- Ensemble various kinds of models to boost the performance.
- Detailed relations between some key factors and accident severity can be further studied and discussed.

Appendix

The python codes and the readme file are provided in the appendix zip file.

References

- https://www.kaggle.com/datasets/sobhanmoosavi/us-accidents
- https://www.kaggle.com/code/jingzongwang/usa-car-accidents-severity-prediction#USA-Car-Accidents-Severity-Prediction
- https://www.kaggle.com/code/deepakdeepu8978/how-severity-the-accidents-is
- https://scikit-learn.org/1.1/