Exploring Factors that Impact Car Accident Severity

Data Mining Final Project Group #5
Abdulaziz Gebril, Jenny Tsai, & Mojahid Osman
April 28, 2020

Introduction

- Car accidents take away people's lives everyday
- US Department of Transportation Stats in 2018:
  - 36,560 deaths
  - 33,654 fatal crashes (severe accidents)

Problem Statement

"What factors might impact <u>car accident severity</u>?"

- Weather conditions
- Road conditions

About the Dataset

- The U.S. accident data are collected from February 2016 to December 2019
- Records gathered using several data providers, including two APIs that provide streaming traffic incident data.
- There are about 3.0 million accident records in this dataset

About the Dataset (cont’d)

49 Features:

- Weather Conditions
  - Temperature
  - Humidity
  - Pressure
  - Visibility
  - Precipitation
  - Wind Chill
  - Wind Speed
  - Weather condition
  - Wind direction
  - Sunrise/Sunset
- Road Conditions:
  - Amenity
  - Bump
  - Crossing
  - Give Way
  - Junction
  - No Exit
  - Railway
  - Roundabout
  - Station
  - Stop
  - Traffic Calming
  - Traffic Signal
- Location / Time
  - State
  - County
  - City
  - Latitude
  - Longitude
  - Start Time
  - End Time

About the Dataset (cont'd)

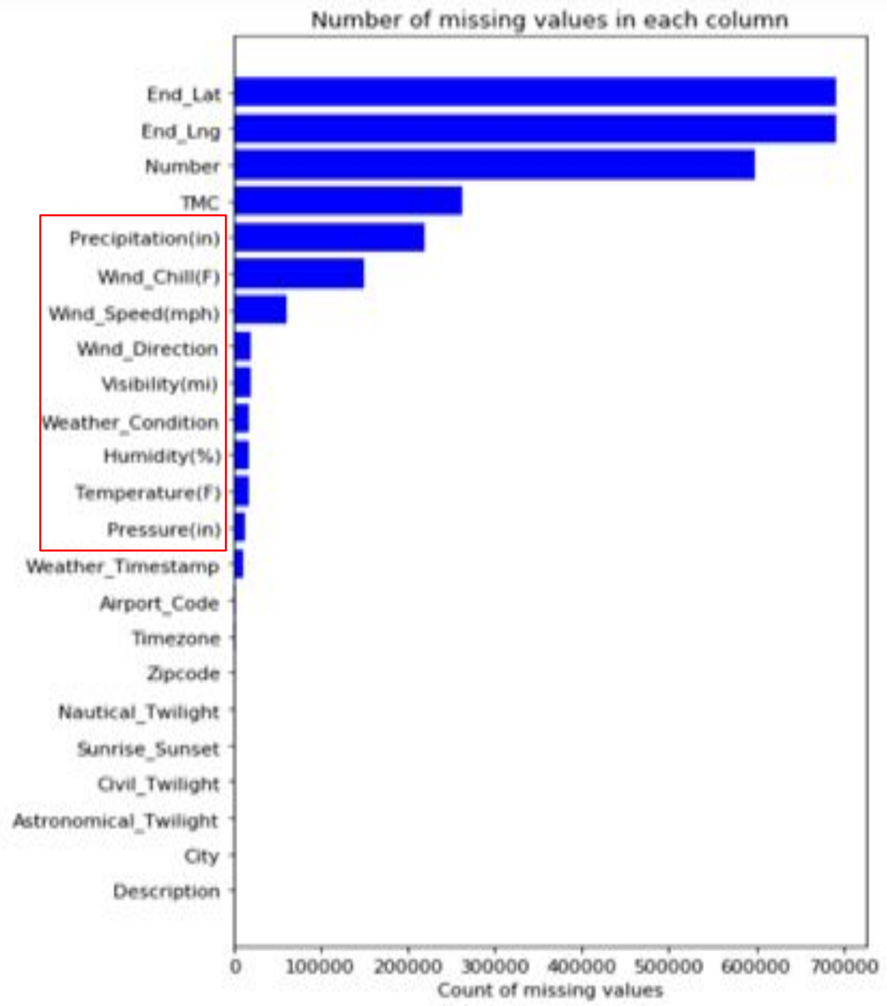
Target Variable:

- Car Accident Severity (1 - 4)
  - Duration of accident

Pre-Processing

- Using Columns **“Start_Time”** and **“End_Time”** to create **“Date”**, **“Year”**,**” Month”**, **”Day”**, **“Hour”**, **“WeekDay”** and **“Time Duration(min)”**

- Severity Classification (High & Low)


Number of missing values in each column
End_Lat
End_Lng
Number
TMC
Precipitation(in)
Wind_Chill(F)
Wind_Speed(mph)
Wind_Direction
Visibility(mi)
Weather_Condition
Humidity(%)
Temperature(F)
Pressure(in)
Weather_Timestamp
Airport_Code
Timezone
Zipcode
Nautical_Twilight
Sunrise_Sunset
Civil_Twilight
Astronomical_Twilight
City
Description
0
100000
200000
300000
400000
500000
600000
700000
Count of missing values

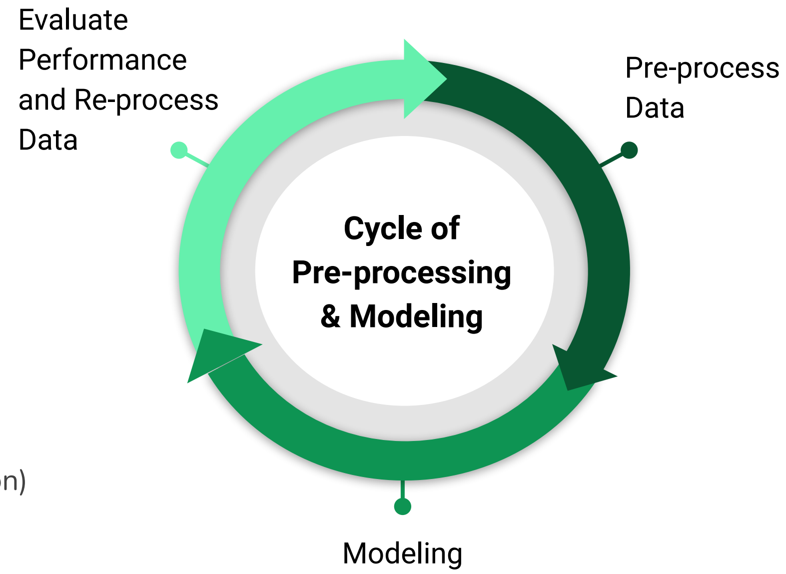
Pre-Processing (cont’d)

- Continuous Variables:
  - (1) Averaging data points that occurred on the same Date and City
  - (2) Averaging data points that occurred on the same Date and State
  - Distance function to check Second Imputation approach.
- Categorical Variables:
  - Collapse categories for weather condition and wind speed
  - Forward fill all data points that occurred on the same Date and City

2^{nd} Stage Pre-processing: For Modeling

An Iterative Process of Cleaning:

- Adjust for imbalance data
  - Subset for 2019 data (whole) and 2018 (high severity)
  - Also tried resampling, but prone to overfitting and underfitting
- Drop attributes not useful to our study
  - Location and Time
  - E.g., Weather condition, Wind Direction, Turning Loop
- Drop attributes still with many nans after imputation
  - Precipitation and Wind Chill (tried regression imputation)



Modeling: Theoretical Framework

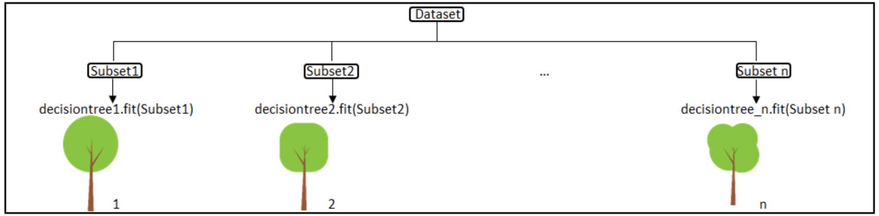
Ensemble Learning

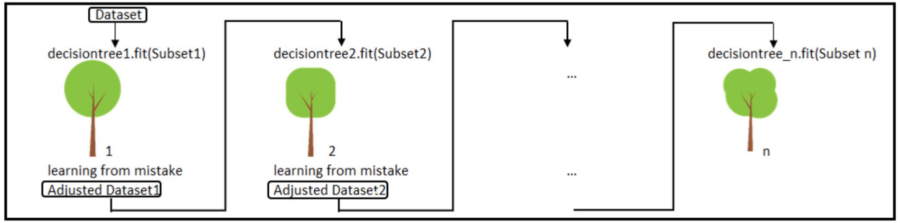
Random Forest - Bagging

- Train a number of trees in parallel
- Make final prediction based on majority vote

Adaptive Boosting

- Train a number of trees in a sequential way
- Learn from previous mistakes and increase the weight of misclassified data points





Modeling: Theoretical Framework (cont'd)

Grid Search + K-Fold Cross Validation

- Find the best hyper-parameters for the model through exhaustive search
- Cross-validated to get reliable results, not just from a particular train-test set

Modeling: Algorithm for Actual Data

sklearn.ensemble

- RandomForestClassifier
- AdaBoostClassifier

sklearn.model_selection

- GridSearch
    - Can specify parameter k for cross-validation

Modeling: Analyses

18 Features:

- 6 Weather Conditions
  - Temperature
  - Humidity
  - Pressure
  - Visibility
  - Wind Speed
  - Sunrise/Sunset

- 12 Road Conditions
  - Amenity
  - Bump
  - Crossing
  - Give Way
  - Junction
  - No Exit
  - Railway
  - Roundabout
  - Station
  - Stop
  - Traffic Calming
  - Traffic Signal

Modeling: Analyses

Steps:

1. Perform Grid Search and CV to find best parameters for RF and AdaBoost using subsample (n = 2,000)
2. Split full dataset (n = 1.2 m) into train & test (7:3)
3. Perform training for RF and AdaBoost using all features

4. Select top 10 important features and re-run the models

5. Build confusion matrix and evaluate model performance against test set

Results: Grid Search & Cross Validation

Grid Search

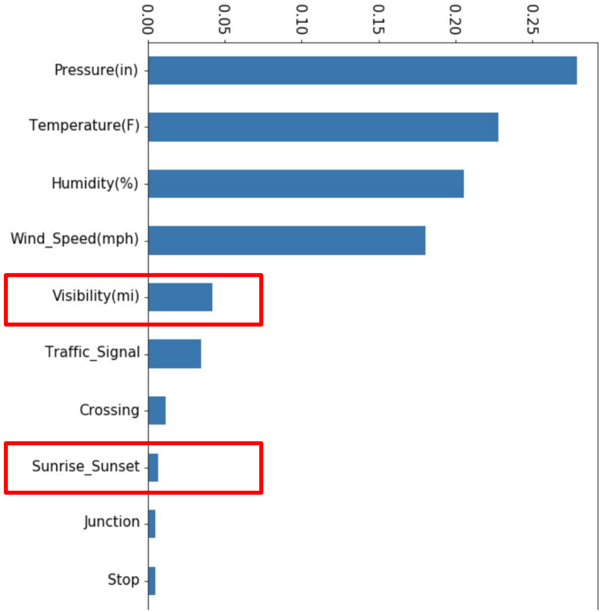
- # of trees, cost function (gini or entropy), learning rate
- Cross validation (k = 5)
- Scoring = Accuracy

Best Parameters for...

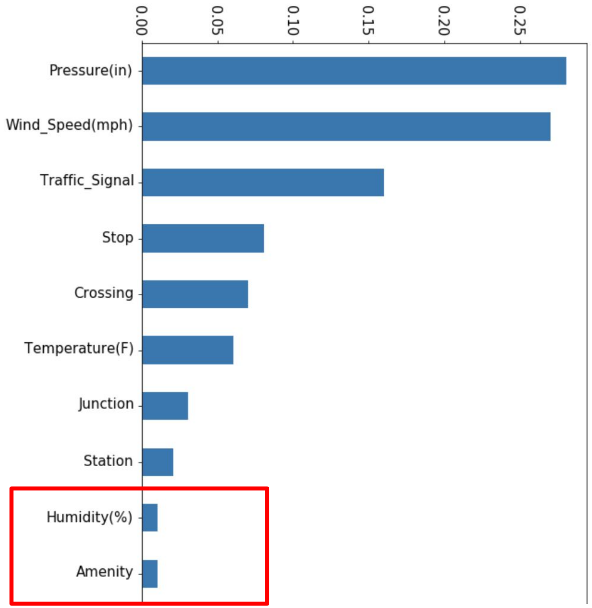
- **Random Forest:**
  - N_estimators = 100
  - Criterion = gini
- **AdaBoost DT**
  - N_estimators = 100
  - Learning_rate = 0.1

Results: Feature Selection

Random Forest



AdaBoost



Results: Feature Selection (cont’d)

Weather Conditions	Road Conditions
• Pressure • Temperature • Humidity • Wind Speed	• Traffic Signal • Crossing • Junction • Stop

Results: Model Performance

Random Forest

```
Results Using Top 10 features:

Classification Report:
              precision    recall  f1-score   support

           0       0.75      0.79      0.77    199174
           1       0.73      0.68      0.70    161896

    accuracy                           0.74    361070
   macro avg       0.74      0.74      0.74    361070
weighted avg       0.74      0.74      0.74    361070



Accuracy :  74.2185725759548


ROC_AUC :  81.16041778890892
```

AdaBoost

```
Results Using Top 10 Features:

Classification Report:
              precision    recall  f1-score   support

           0       0.70      0.73      0.72    199174
           1       0.65      0.62      0.64    161896

    accuracy                           0.68    361070
   macro avg       0.68      0.68      0.68    361070
weighted avg       0.68      0.68      0.68    361070



Accuracy :  68.25546292962584


ROC_AUC :  73.30263509189082
```

Summary & Conclusion

- 8 features that impact accident severity in both models:
    - Pressure
    - Temp
    - Humidity
    - Wind Speed
    - Traffic Signal
    - Crossing
    - Junction
    - Stop

- RF favors weather variables, while AdaBoost favors road variables

- Overall, data fits better with RF
    - Why?

Visualization: PyQT5