Predicting Car Accident Severity

to improve public safety outreach in the United States

Team 4: Thomas Berdahl and Alyssa Duran
DAAN 888 Spring 2024

Project Research Question

Background: In 2021, there were 39,508 fatal motor vehicle crashes accounting for 42,939 deaths in the United States(Fatality Facts, n.d.). This is equal to 12.9 deaths per 100,000 people, and 1.37 deaths per 100 million miles traveled.

Scope: We utilized US car accident data to predict accident severity, considering factors like weather, time, and road conditions. Our model aimed to help government agencies issue real-time alerts about hazardous conditions, enhancing driver awareness and safety.

.

Dataset Description and Source

Description: This dataset contains about 7.7 million accident records from February 2016 to March of 2023 in 49 states of the USA. Attributes include time, location, street conditions, weather, light conditions, and accident severity. With this dataset we will be able to evaluate the components and build a supervised predictive model.

Source: Kaggle

Data Exploration and Issues

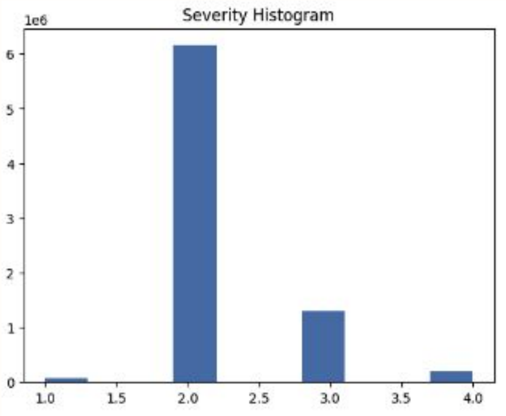
Columns	46
Rows	Approx 7.7M

Attribute	Data type	Measure Type	Attribute	Data type	Measure Type
Airport_Code	String	Nominal	**Roundabout**	Boolean	Nominal
Amenity	Boolean	Nominal	**Severity**	Integer	Ordinal
Astronomical_Twiligh	String	Nominal	**Source**	String	Nominal
Bump	Boolean	Nominal	**Start_Lat**	Decimal	Interval
City	String	Nominal	**Start_Lng**	Decimal	Interval
Civil_Twilight	String	Nominal	**Start_Time**	Datetime	Ordinal
Country	String	Nominal	**State**	String	Nominal
County	String	Nominal	**Station**	Boolean	Nominal
Crossing	Boolean	Nominal	**Stop**	Boolean	Nominal
Description	String	Nominal	**Street**	String	Nominal
Distance(mi)	Decimal	Ratio	**Sunrise_Sunset**	String	Nominal
End_Lat	Decimal	Interval	**Temperature(F)**	Decimal	Interval
End_Lng	Decimal	Interval	**Timezone**	String	Nominal
End_Time	Datetime	Ordinal	**Traffic_Calming**	Boolean	Nominal
Give_Way	Boolean	Nominal	**Traffic_Signal**	Boolean	Nominal
Humidity(%)	Integer	Ratio	**Turning_Loop**	Boolean	Nominal
ID	String	Nominal	**Visibility(mi)**	Decimal	Ratio
Junction	Boolean	Nominal	**Weather_Condition**	String	Nominal
Nautical_Twilight	String	Nominal	**Weather_Timestam**	Datetime	Ordinal
No_Exit	Boolean	Nominal	**Wind_Chill(F)**	Decimal	Interval
Precipitation(in)	Decimal	Ratio	**Wind_Direction**	String	Nominal
Pressure(in)	Decimal	Ratio	**Wind_Speed(mph)**	Decimal	Ratio
Railway	Boolean	Nominal	**Zipcode**	String	Nominal

Data Exploration and Issues (Continued)

Severity (Response Variable) Distribution

Value	Distribution	Meaning
1	0.44% (67K)	Lowest Severity; No Delay
2	79.97% (6.15M)	Minimal Severity & Delays
3	16.86% (1.29M)	Severe; Significant Delays
4	2.66% (205K)	Highest Severity; Extreme Delays



- Data Imbalance of response variable
  - Sampling methods will need to be used

Data Exploration and Issues (Continued)

- Columns with Low Variance:
  - Distance(mi)
  - Turning_Loop - 100% False
- Columns with No Relevance:
  - Wind_Chill(F) - Highly correlated with temperature (r=0.99)
  - 16 Other Columns
    - ID, State, County, Airport_Code, etc.

Data Cleaning

- Variable Transformation
  - Boolean variables from 0 to 1 (16 Columns)
  - Split date & time into separate columns
    - Date: 2016-02-08 07:44:26
    -

Severity	Year	Month	Hour_of_Day	Day_of_Week
1	2016	2	7	0

  - String to number transformations using dictionary
    - Values in 'Weather_Conditions' were aggregated with a dict, split on comma, and binarized.
      Example:

SEVERITY	WIND	RAIN	SNOW/ICE	CLEAR	FOG	CLOUDY	THUNDER	TORNADO	DUSTSTORM /DEBRIS
4	1	0	1	0	0	1	0	0	0

Data Cleaning

- Variable Transformation
  - Boolean variables from 0 to 1 (16 Columns)
  - Split date & time into separate columns
    - Date: 2016-02-08 07:44:26
Severity	Year	Month	Hour_of_Day	Day_of_Week
1	2016	2	7	0
  - String to number transformations using dictionary
    - Values in ‘Weather_Conditions’ were aggregated with a dict, split on comma, and binarized.
      Example:

SEVERITY	WIND	RAIN	SNOW/ICE	CLEAR	FOG	CLOUDY	THUNDER	TORNADO	DUSTSTORM /DEBRIS
4	1	0	1	0	0	1	0	0	0

Data Cleaning (Continued)

- Remove Null Values
- Outlier Detection & Removal
  - IQR Formula
  - Performed on ‘Temperature(F)’, Pressure(in)’, ‘Wind_Speed(mph)’.

- Cleaning Summary
  - Removed 2.93M values; Left with 4790161

Model Preparation

- Multicollinearity
- Data Partitioning
- Data Imbalance
  - Random Undersampling
  - Oversampling using SMOTE
  - Random Forest Parameter class_weight =’balanced’

Training data set distribution:

‘Severity’ Value	Total Values	Distribution PCT
1 (Least Severe)	43,552	1.1%
2	3,237,561	84.5%
3	462,597	12.1%
4 (Most Severe)	88,418	2.3%

Model Parameters

- Model Type: Random Forest
- Training/Testing/Validation Split: 80/10/10
- n_estimators: Grid Search (Use k-fold validation)
- max_depth: Grid Search (Use k-fold validation)
- max_features: Grid Search (Use k-fold validation)
- min_samples_leaf: Grid Search (Use k-fold validation)
- min_samples_split: Grid Search (Use k-fold validation)

Hyperparameter Tuning

Grid search completed on undersampled balanced training data set with 5-fold cross validation

Grid #	max_depth	max_features	n_estimators	min_samples_leaf	min_samples_split	Accuracy
Grid 1	5	5	100	2 (default)	1 (default)	0.4375
Grid 2	None	5	200	2 (default)	1 (default)	0.5358
Grid 3	None	sqrt	50	2 (default)	1 (default)	0.5273
Grid 4	None	sqrt	50	1	4	0.5293

Final Model Parameters

- Model Type: Random Forest
- Training/Testing/Validation Split: 80/10/10
- n_estimators: 50
- max_depth: None
- max_features: ‘sqrt’
- min_samples_leaf: 2 (default)
- min_samples_split: 1 (default)
- class_weight='balanced' (to address data imbalance)

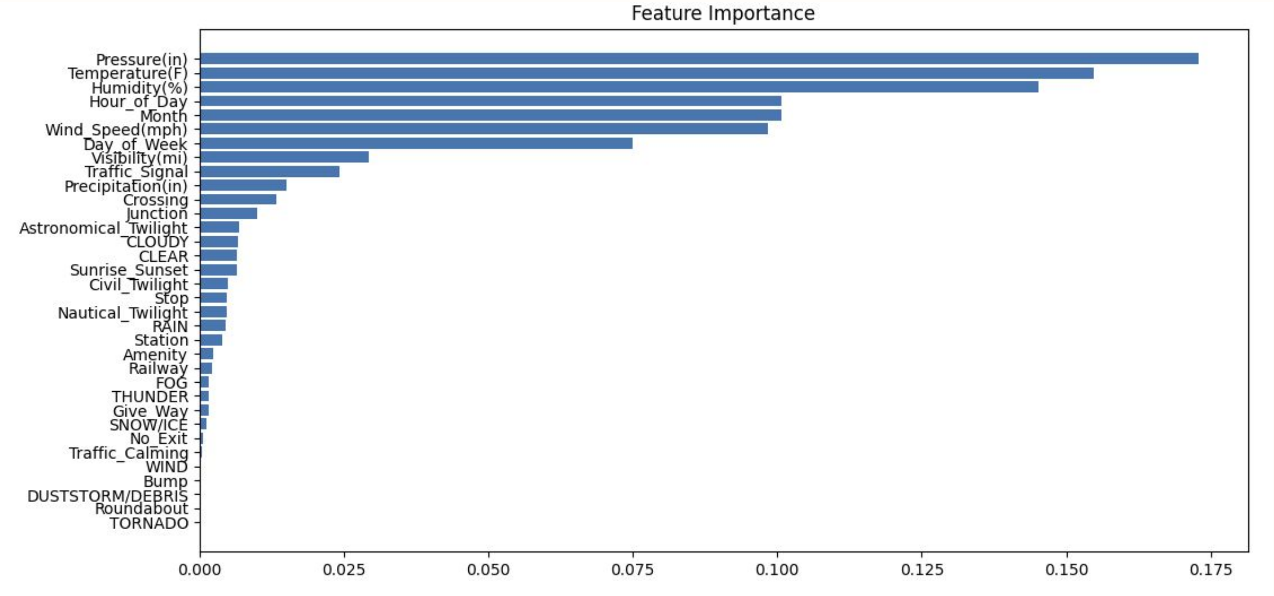
Model Performance

- Test set accuracy of 82.84 %
- Validation set accuracy of 82.91%

Outcome	Precision	Recall	F1-score	Support
1	0.50	0.22	0.30	5407
2	0.86	0.95	0.90	404588
3	0.41	0.14	0.21	57972
4	0.29	0.31	0.30	11049

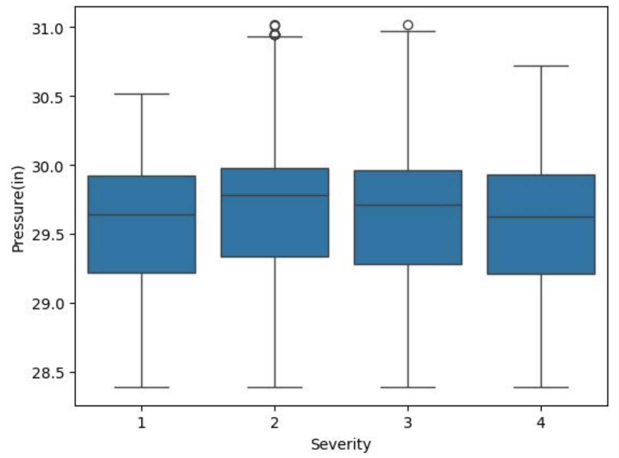
Variable Importance

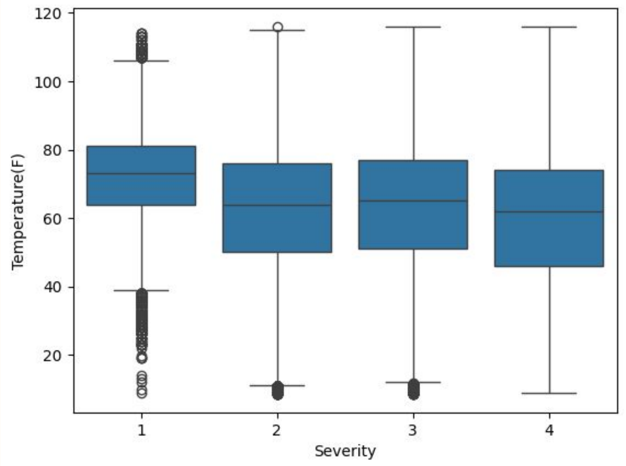
Variable	Importance
Pressure(in)	0.17
Temperature (F)	0.15
Humidity(%)	0.15
Wind_Speed(mph)	0.1
Month	0.1
Hour_of_Day	0.1



Findings - Pressure and Temperature

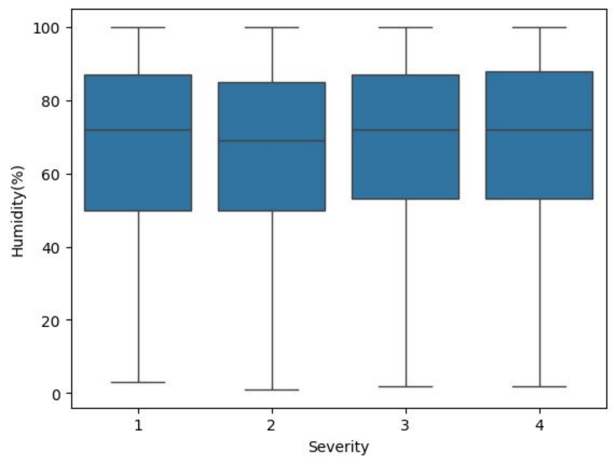
- Lower Pressure and Temperatures had more severe accidents



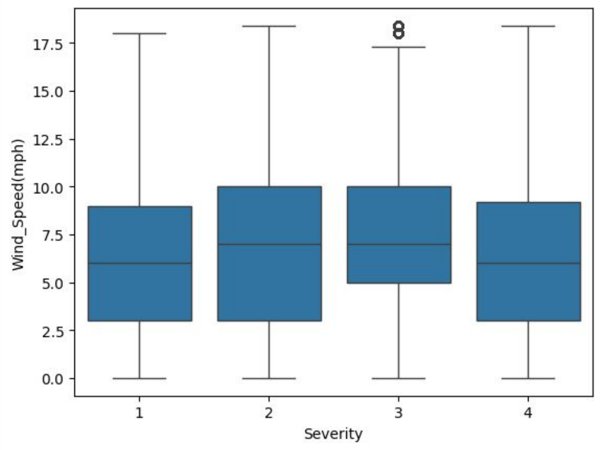


Findings - Humidity and Wind Speed

- Higher humidity had more severe accidents



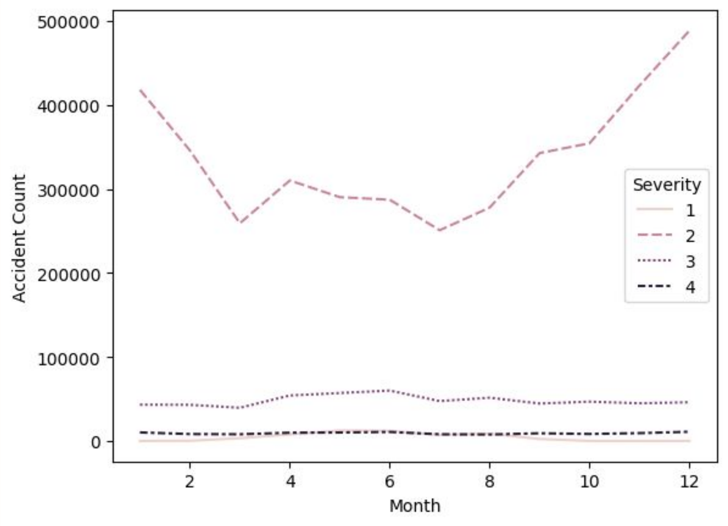
- Accidents with a severity of 3 have a higher median wind speed

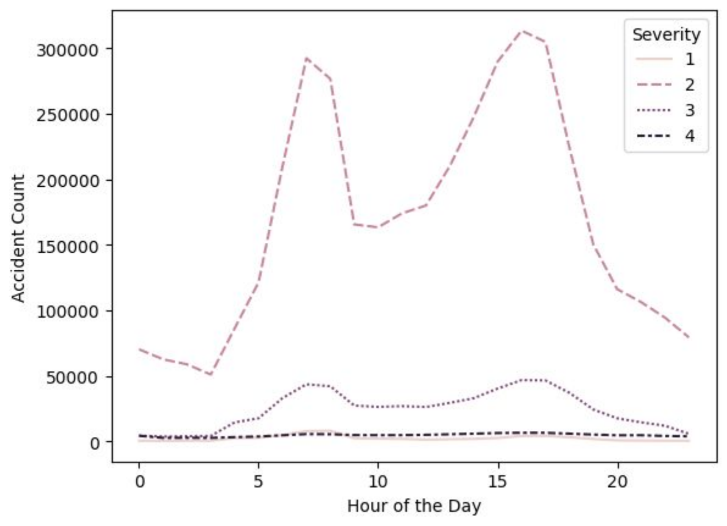


Findings - Month and Hour

- High frequency in severity 2 accidents in the months of November through January
- Severity 3 accidents increase around April to June

- Accidents with a severity of 2 and 3 occur most often during morning and afternoon commute hours





Conclusion

- Potential for the predictive model to utilize real-time weather measurements like pressure, temperature, humidity, and wind speed to issue alerts to the public of hazardous conditions
- Media campaigns during high-accident risk times can help drivers be mindful of their driving
  - Safe driving messages on the morning news
  - PSA during winter with winter driving tips
- Challenges with predicting the most severe accidents but benefit of reducing overall accident frequencies

Limitations

- The computational requirements for our large dataset were not met, resulting in extensive code execution times. The follow limitations were introduced:
  - Reduced performance of hyperparameter tuning using grid search
  - Oversampling methods (SMOTE) to address data imbalance were unsuccessful.
    - Undersampling was used as an alternative, and likely removed variance.
      - Severity ‘2’ value counts went from 3,624,973 to 49,669.
- Accident severity is difficult to predict. Pulling additional data from another source may have introduced more significant variables, contributing to a better performing model.
  - Example: zip code aggregated demographic data