UNIVERSITY OF NORTH CAROLINA
CHARLOTTE

PREDICTING TRAFFIC IMPACT OF ACCIDENTS

A Machine Learning Classifier Based Approach to Integrating Location, Temporal, and Environmental Factors in Estimating Traffic Disruptions

Michael Tricanowicz – DSBA-6156 – Fall 2024

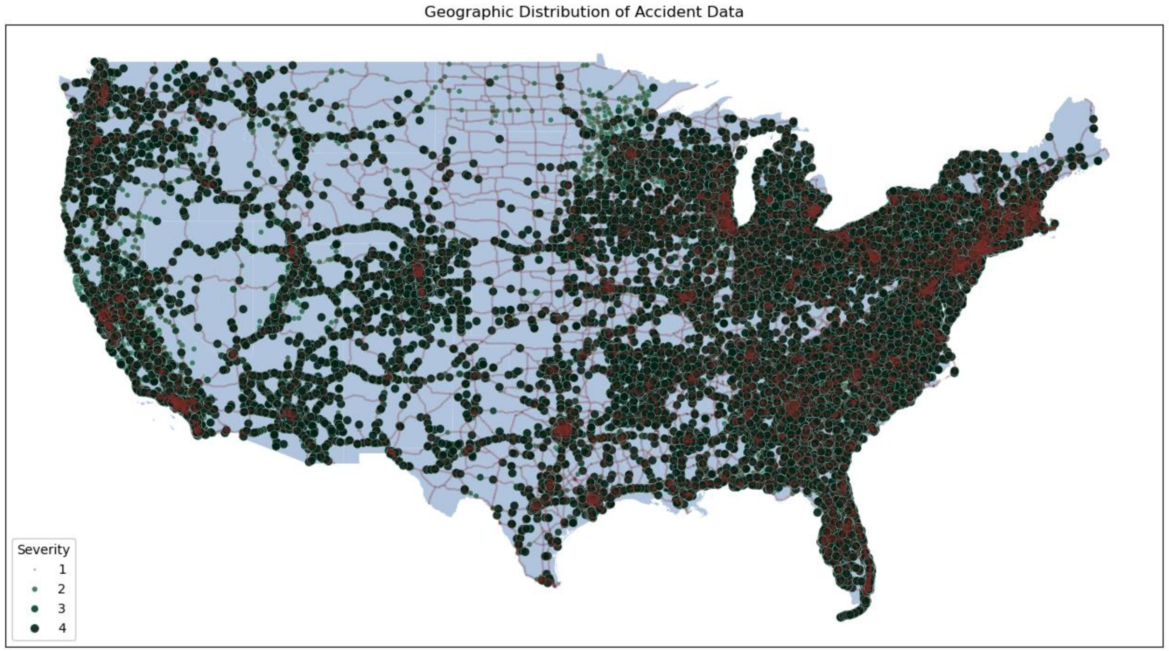
Overview

- Review the data set used
- Explain data cleaning process
- Discuss model development and refinement
- Demo predictor app

The Data

- Analysis dataset of approximately 270k entries of car accidents in the contiguous United States that occurred between 2016 and 2023.
- Data was undersampled from original set with more than 7.7 million entries.
- **Target feature: Severity.** Severity is rated on a scale of 1 (least severe) to 4 (most severe). Undersampled set has an equal number of entries for each severity.
- There are 45 independent features.

COLUMN	DESCRIPTION	TYPE
ID	This is a unique identifier of the accident record	object
Source	Source of raw accident data	object
Severity	Shows the severity of the accident, a number between 1 and 4, where 1 indicates the least impact on traffic (i.e., short delay)	int
Start_Time	Shows start time of the accident in local time zone.	object
End_Time	Shows end time of the accident in local time zone. End time here refers to when the impact of accident on traffic flow	object
Start_Lat	Shows latitude in GPS coordinate of the start point.	float
Start_Lng	Shows longitude in GPS coordinate of the start point.	float
End_Lat	Shows latitude in GPS coordinate of the end point.	float
End_Lng	Shows longitude in GPS coordinate of the end point.	float
Distance(mi)	The length of the road extent affected by the accident in miles.	float
Description	Shows a human provided description of the accident.	object
Street	Shows the street name in address field.	object
City	Shows the city in address field.	object
County	Shows the county in address field.	object
State	Shows the state in address field.	object
Zipcode	Shows the zipcode in address field.	object
Country	Shows the country in address field.	object
Timezone	Shows timezone based on the location of the accident (eastern, central, etc.).	object
Airport_Code	Denotes an airport-based weather station which is the closest one to location of the accident.	object
Weather_Timestamp	Shows the time-stamp of weather observation record (in local time).	object
Temperature(F)	Shows the temperature (in Fahrenheit).	float
Wind_Chill(F)	Shows the wind chill (in Fahrenheit).	float
Humidity(%)	Shows the humidity (in percentage).	float
Pressure(in)	Shows the air pressure (in inches).	float
Visibility(mi)	Shows visibility (in miles).	float
Wind_Direction	Shows wind direction.	object
Wind_Speed(mph)	Shows wind speed (in miles per hour).	float
Precipitation(in)	Shows precipitation amount in inches, if there is any.	float
Weather_Condition	Shows the weather condition (rain, snow, thunderstorm, fog, etc.)	object
Amenity	A POI annotation which indicates presence of amenity in a nearby location.	bool
Bump	A POI annotation which indicates presence of speed bump or hump in a nearby location.	bool
Crossing	A POI annotation which indicates presence of crossing in a nearby location.	bool
Give_Way	A POI annotation which indicates presence of give_way in a nearby location.	bool
Junction	A POI annotation which indicates presence of junction in a nearby location.	bool
No_Exit	A POI annotation which indicates presence of no_exit in a nearby location.	bool
Railway	A POI annotation which indicates presence of railway in a nearby location.	bool
Roundabout	A POI annotation which indicates presence of roundabout in a nearby location.	bool
Station	A POI annotation which indicates presence of station in a nearby location.	bool
Stop	A POI annotation which indicates presence of stop in a nearby location.	bool
Traffic_Calming	A POI annotation which indicates presence of traffic_calming in a nearby location.	bool
Traffic_Signal	A POI annotation which indicates presence of traffic_signal in a nearby location.	bool
Turning_Loop	A POI annotation which indicates presence of turning_loop in a nearby location.	bool
Sunrise_Sunset	Shows the period of day (i.e. day or night) based on sunrise/sunset.	object
Civil_Twilight	Shows the period of day (i.e. day or night) based on civil twilight.	object
Nautical_Twilight	Shows the period of day (i.e. day or night) based on nautical twilight.	object
Astronomical_Twilight	Shows the period of day (i.e. day or night) based on astronomical twilight.	object



The goal is to predict the ultimate severity of traffic impact *at the start* of an accident event.

Understanding the Data

- High dimensionality in dataset
- Significant number of missing values
- Extreme cardinality in categorical and descriptive features
- Uneven representation of Boolean features
- Unnecessary features

Initial Feature Set

	Feature	Number Unique	Percent Null
0	ID	269464	0.00
1	Source	3	0.00
2	Severity	4	0.00
3	Start_Time	255226	0.00
4	End_Time	258776	0.00
5	Start_Lat	202556	0.00
6	Start_Lng	203010	0.00
7	End_Lat	113824	47.14
8	End_Lng	114176	47.14
9	Distance(mi)	9473	0.00
10	Description	224436	0.00
11	Street	60425	0.16
12	City	9386	0.00
13	County	1631	0.00
14	State	49	0.00
15	Zipcode	77382	0.03
16	Country	1	0.00
17	Timezone	4	0.10
18	Airport_Code	1904	0.33
19	Weather_Timestamp	158071	1.60
20	Temperature(F)	673	2.26
21	Wind_Chill(F)	768	25.06
22	Humidity(%)	100	2.42
23	Pressure(in)	965	1.90
24	Visibility(mi)	54	2.40
25	Wind_Direction	24	2.30
26	Wind_Speed(mph)	86	7.59
27	Precipitation(in)	150	28.36
28	Weather_Condition	93	2.39
29	Amenity	2	0.00
30	Bump	2	0.00
31	Crossing	2	0.00
32	Give_Way	2	0.00
33	Junction	2	0.00
34	No_Exit	2	0.00
35	Railway	2	0.00
36	Roundabout	2	0.00
37	Station	2	0.00
38	Stop	2	0.00
39	Traffic_Calming	2	0.00
40	Traffic_Signal	2	0.00
41	Turning_Loop	1	0.00
42	Sunrise_Sunset	2	0.44
43	Civil_Twilight	2	0.44
44	Nautical_Twilight	2	0.44
45	Astronomical_Twilight	2	0.44

→ ?

Cleaning the Data

Drop columns

- Source: Unnecessary.
- End_Lat, End_Lng: High nulls, no unique signal.
- Bump, Give_Way, No_Exit, Railway, Roundabout, Traffic_Calming, Turning_Loop: Boolean features >99% false. Entries that are true are nearly evenly split between severities. No useful signal.
- Weather_Condition: High cardinality (93 unique values). Unneeded subjective categorical feature given the presence of other objective weather features.

	True	% True	False	% False
Amenity	3210.0	1.19	266254.0	98.81
Bump	84.0	0.03	269380.0	99.97
Crossing	34129.0	12.67	235335.0	87.33
Give_Way	1541.0	0.57	267923.0	99.43
Junction	21343.0	7.92	248121.0	92.08
No_Exit	790.0	0.29	268674.0	99.71
Railway	2547.0	0.95	266917.0	99.05
Roundabout	5.0	0.00	269459.0	100.00
Station	6248.0	2.32	263216.0	97.68
Stop	7029.0	2.61	262435.0	97.39
Traffic_Calming	224.0	0.08	269240.0	99.92
Traffic_Signal	48469.0	17.99	220995.0	82.01
Turning_Loop	0.0	0.00	269464.0	100.00

Drop rows

- All rows with Weather_Timestamp = null. These rows also have null values for ALL weather features making imputation impractical. Represents 1.60% of total entries.
- All rows with null values in all day/night features. Represents 0.44% of total entries.
- All rows with Street = null. Represents 0.16% of total entries.
- Dropping these rows also reduce null values in the other weather features.

Add features

- Extract the components of the start time: hour, day of week, month, year. ML models cannot parse datetime data type.

Imputing Missing Data

Wind_Direction, Temperature(F), Humidity(%), Pressure(in), Visibility(mi), Wind_Speed(mph)

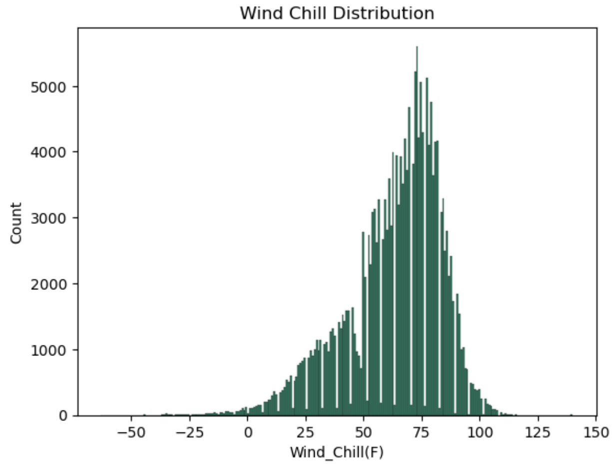
- Single digit percent null (most less than 1%)
- Impute missing values by using the mean of the feature grouped by Month and Zipcode then Month and *State*

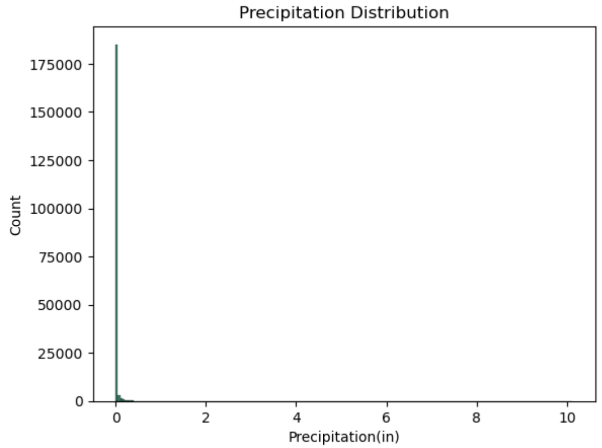
Wind_Chill(F)

- 24% null
- Impute missing values using a linear regression model trained on the set of non-null wind chill values.
- Independent features are the other weather features above (nulls already removed) plus Zipcode and Month
- Trained model accuracy is 99% on the test data. Deemed acceptable to apply to impute the missing values.

Precipitation(in)

- 27% null
- Of non-null entries, 92% are 0 inches and 4% are less than 0.05 inches
- Attempted to setup a two model imputation procedure: logistic regression to predict zero or nonzero and linear regression to predict the value for a predicted nonzero.
- Independent features are the other weather features above (nulls already removed) plus Zipcode and Month
- Trained logistic regression model accuracy is 92%, but linear regression model is 15%. Deemed not acceptable.
- Because the data is so skewed to zero or trace amounts of precipitation, null values are uniformly set to 0 rather than further pursue the complex model setup.





Preparing the Data

- High dimensionality remains despite reduction. Additional features to be dropped during model development.
- Missing values eliminated
- Extreme cardinality in categorical and descriptive features remains despite reduction
- Uneven representation of Boolean features addressed
- Unnecessary features removed

Initial Feature Set

	Feature	Number Unique	Percent Null
0	ID	269464	0.00
1	Source	3	0.00
2	Severity	4	0.00
3	Start_Time	255226	0.00
4	End_Time	258776	0.00
5	Start_Lat	202556	0.00
6	Start_Lng	203010	0.00
7	End_Lat	113824	47.14
8	End_Lng	114176	47.14
9	Distance(mi)	9473	0.00
10	Description	224436	0.00
11	Street	60425	0.16
12	City	9386	0.00
13	County	1631	0.00
14	State	49	0.00
15	Zipcode	77382	0.03
16	Country	1	0.00
17	Timezone	4	0.10
18	Airport_Code	1904	0.33
19	Weather_Timestamp	158071	1.60
20	Temperature(F)	673	2.26
21	Wind_Chill(F)	768	25.06
22	Humidity(%)	100	2.42
23	Pressure(in)	965	1.90
24	Visibility(mi)	54	2.40
25	Wind_Direction	24	2.30
26	Wind_Speed(mph)	86	7.59
27	Precipitation(in)	150	28.36
28	Weather_Condition	93	2.39
29	Amenity	2	0.00
30	Bump	2	0.00
31	Crossing	2	0.00
32	Give_Way	2	0.00
33	Junction	2	0.00
34	No_Exit	2	0.00
35	Railway	2	0.00
36	Roundabout	2	0.00
37	Station	2	0.00
38	Stop	2	0.00
39	Traffic_Calming	2	0.00
40	Traffic_Signal	2	0.00
41	Turning_Loop	1	0.00
42	Sunrise_Sunset	2	0.44
43	Civil_Twilight	2	0.44
44	Nautical_Twilight	2	0.44
45	Astronomical_Twilight	2	0.44

Prepared Feature Set

	Feature	Number Unique	Percent Null
0	ID	263682	0.0
1	Severity	4	0.0
2	Start_Time	248705	0.0
3	End_Time	252441	0.0
4	Start_Lat	197945	0.0
5	Start_Lng	198407	0.0
6	Distance(mi)	9342	0.0
7	Description	219642	0.0
8	Street	59234	0.0
9	City	8907	0.0
10	County	1577	0.0
11	State	49	0.0
12	Zipcode	16316	0.0
13	Country	1	0.0
14	Timezone	4	0.0
15	Airport_Code	1868	0.0
16	Weather_Timestamp	157482	0.0
17	Temperature(F)	1286	0.0
18	Wind_Chill(F)	61945	0.0
19	Humidity(%)	755	0.0
20	Pressure(in)	1329	0.0
21	Visibility(mi)	401	0.0
22	Wind_Direction	18	0.0
23	Wind_Speed(mph)	3322	0.0
24	Precipitation(in)	150	0.0
25	Amenity	2	0.0
26	Crossing	2	0.0
27	Junction	2	0.0
28	Station	2	0.0
29	Stop	2	0.0
30	Traffic_Signal	2	0.0
31	Sunrise_Sunset	2	0.0
32	Civil_Twilight	2	0.0
33	Nautical_Twilight	2	0.0
34	Astronomical_Twilight	2	0.0
35	Start_Year	8	0.0
36	Start_Month	12	0.0
37	Start_Day	7	0.0
38	Start_Hour	24	0.0

The Data…Cleaned and Complete

Entries	Features	Null Values	Severity Distribution
269,464	46	466,044	1 | 67366 | 25% 2 | 67366 | 25% 3 | 67366 | 25% 4 | 67366 | 25%

Drop irrelevant and unnecessary columns.
Correct data types.
Consolidate categorical data.
Impute missing data.
Drop rows that cannot be imputed.

263,682 2% reduction	39 15% reduction	0 100% reduction	1 | 66398 | 25.18% 2 | 66066 | 25.05% 3 | 66388 | 25.18% 4 | 64830 | 24.59%

Modeling

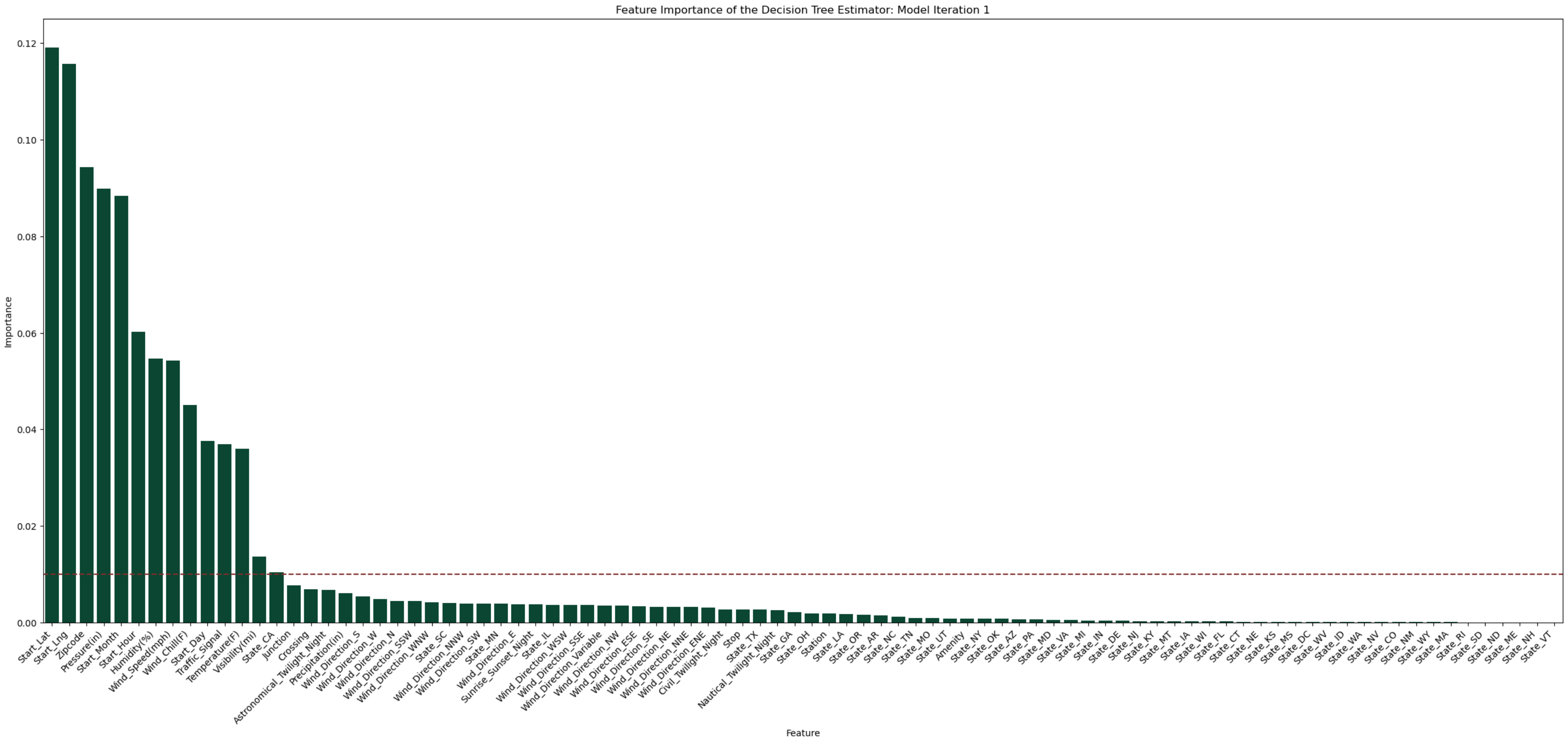
Multiclass classification problem. Tree based modeling methods chosen, starting with a decision tree.

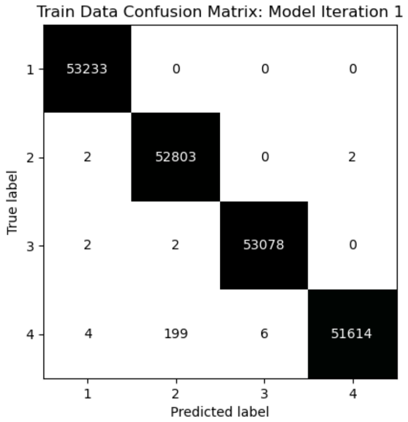
Additional data preparation for modeling:

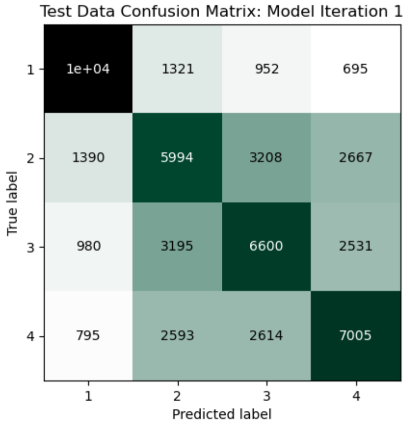
- Drop additional features:
    - ID: Unique identifier. Does not provide any descriptive value.
    - Start_Time, End_Time, Weather_Timestamp: datetime data type cannot be parsed by the model
    - Distance, End_Time: Unknown until after completion of accident event. Not useful for predictive purposes.
    - Description: Long form text, unusable by the model
    - Country: This dataset is entirely from the United States. All entries have the same country, so this feature is not useful.
    - Street, City, County, Timezone, Airport_Code: Categorical features with very, very high cardinality. Would unnecessarily explode dimensionality of data set without providing unique information.
    - Start_Year: Including the year will make future predictions in different years less accurate
- One hot encode categorical and Boolean features

First Attempt

Unconstrained decision tree. Overly complex and overfit.







Improving the Model

Drop additional features:

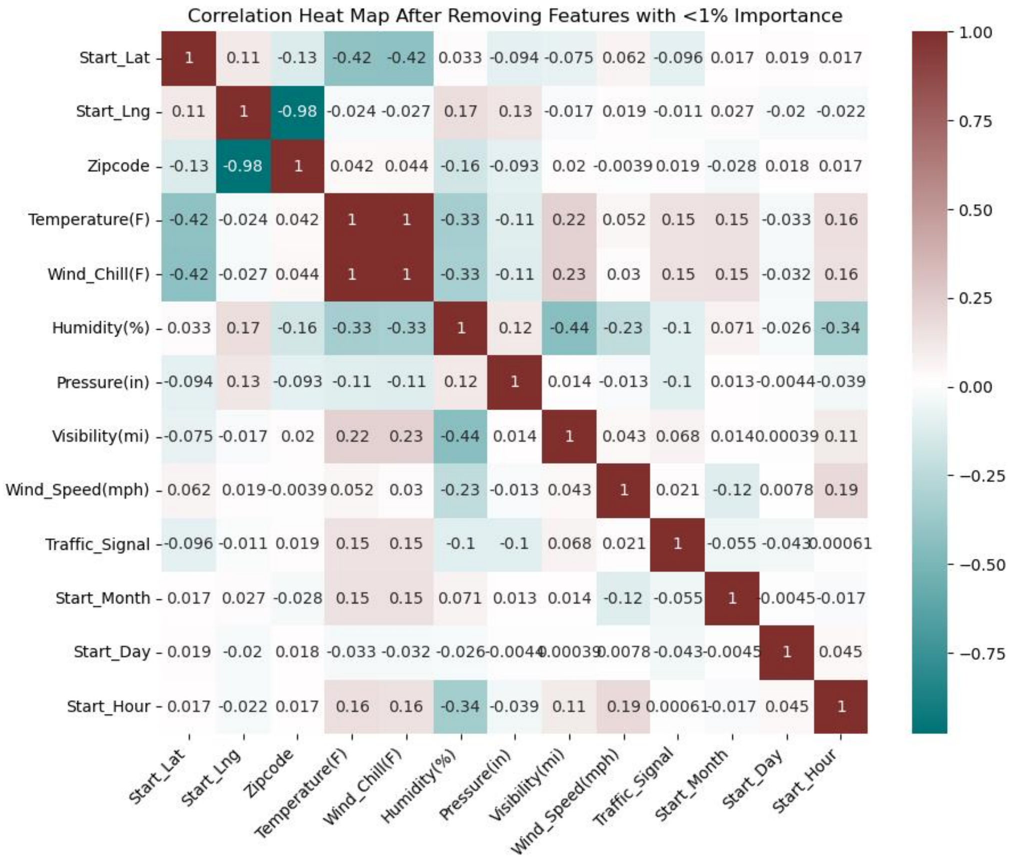
- State, Wind_Direction: Categorical features add little value but drastically increase dimensionality. Start_Lat and Start_Lng features cover location. Wind_Direction has little importance in the model predictions.
- Remaining numeric and Boolean features with <1% feature importance.
- Zipcode: Highly correlated with Longitude (and to a lesser extent Latitude). The Lat/Lon coordinates will adequately provide whatever predictive signal comes from location.
- Wind_Chill(F): 100% correlated with temperature and does not add any additional signal.
- Results in a final model input set of 11 features

Use a Random Forest model:

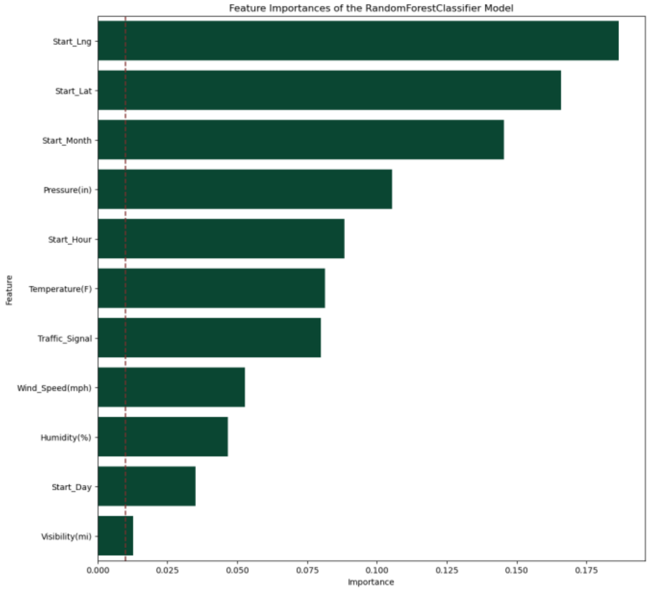
- Attempt to reduce overfitting.
- Constrain and optimize the model using HalvingGridSearchCV

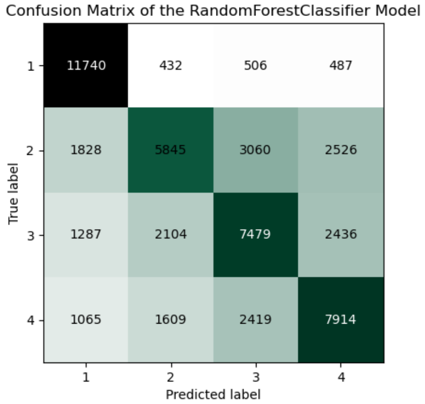
Use an XGBoost model:

- Evaluate for performance differences on same input features and dataset
- Constrain and optimize the model using HalvingGridSearchCV



Preliminary Results

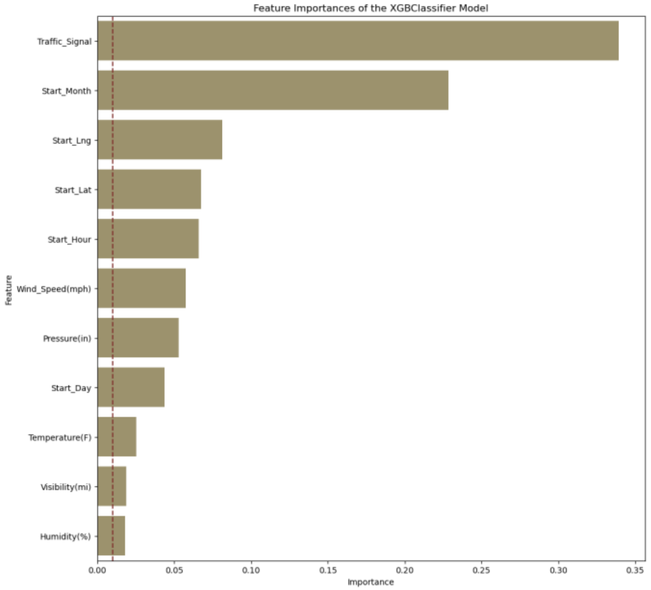


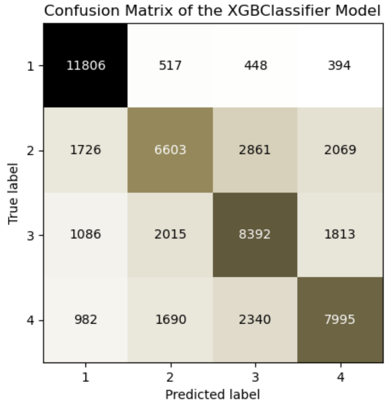


Random Forest
- Train Accuracy = 74%
- Test Accuracy = 63%
- F1 Score = 62% (81%, 50% , 56%, 60%)
- Importance more evenly distributed among input features.
- ~300 MB model file size

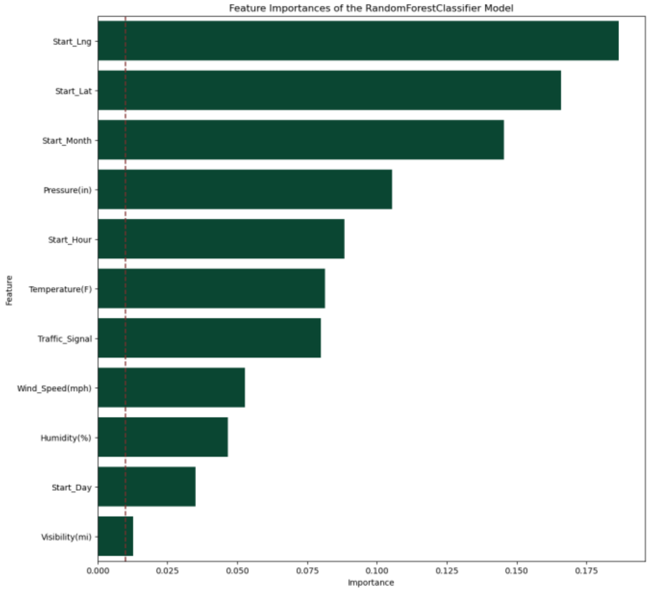
XGBoost
- Train Accuracy = 72%
- Test Accuracy = 66%
- F1 Score = 65% (82%, 55%, 61%, 63%)
- Importance heavily weighted to top 2 features.
- ~2 MB model file size

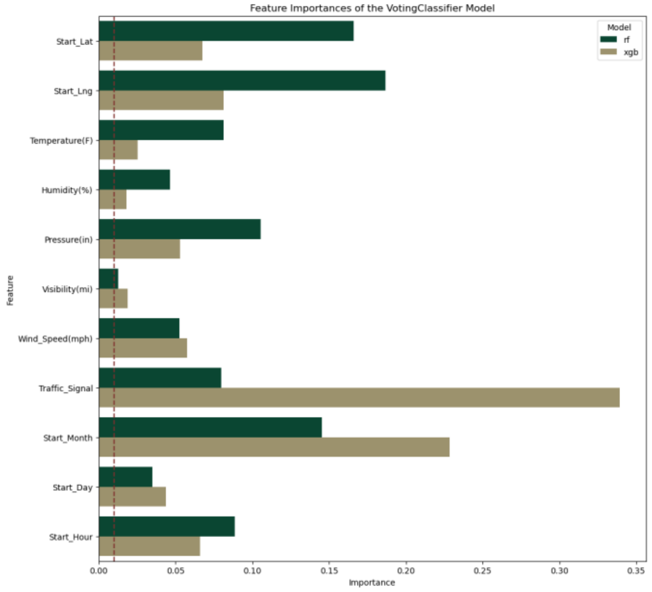
What if the two models were combined?
- Takes advantage of both model structures
- Smooths out variations in predictions due to feature importances
- Accomplished using a Voting Classifier

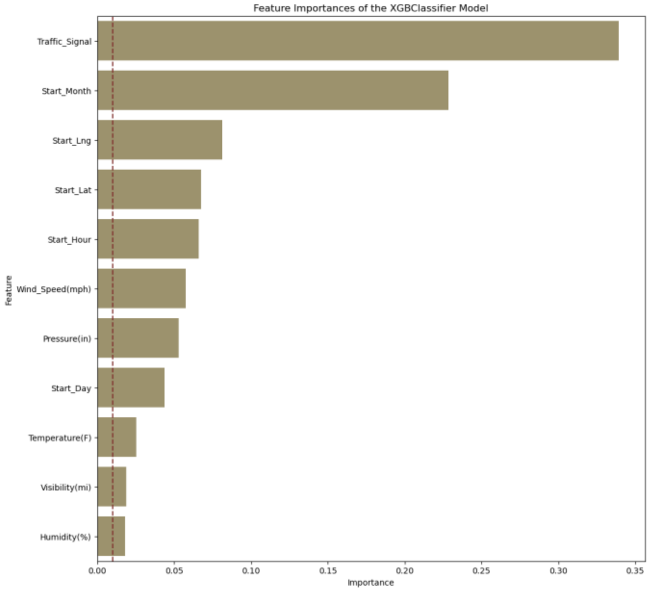


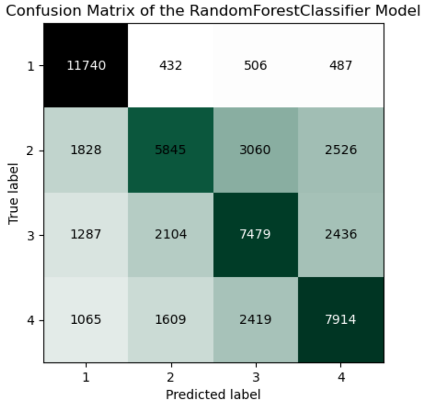


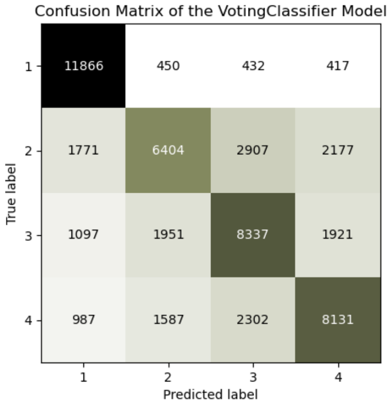
Blending the Models

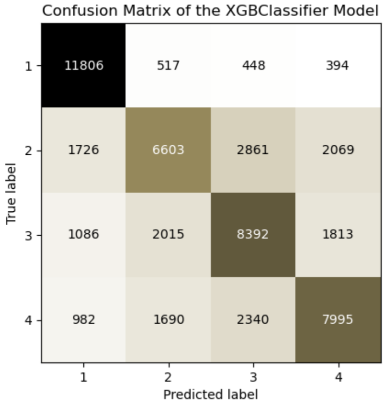






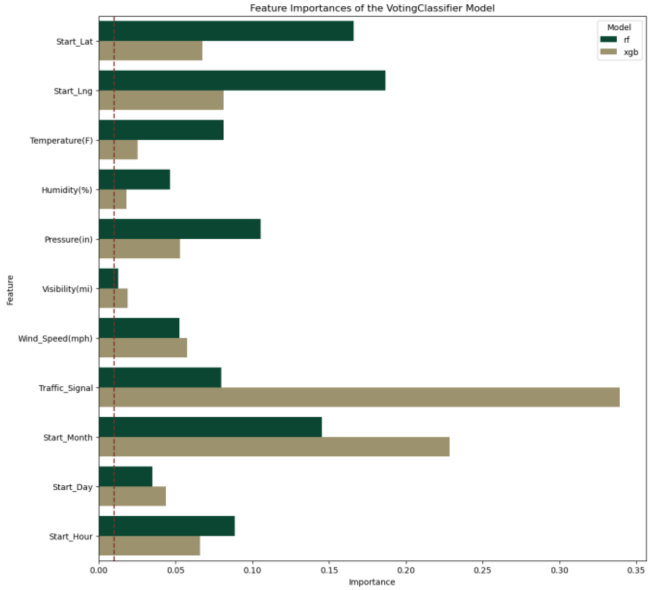


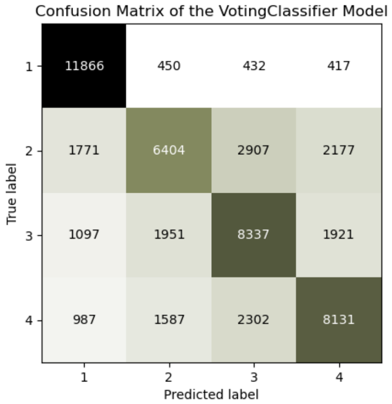




Final Model

- Blended Voting Classifier that combines the optimized Random Forest and XGBoost models
- Soft voting classifier averages the predicted probabilities of the models to determine a class prediction
- Maintains the performance improvements of the XGBoost model
- Evens out feature importance imbalance in the overall prediction process
- Significantly improved predictive accuracy (66%) over random guess (25%)
- ~600 MB model file size





Input Features:

- Latitude
- Longitude
- Temperature
- Humidity
- Barometric Pressure
- Visibility
- Wind Speed
- Presence of Traffic Signal
- Month
- Day of Week
- Hour of Day

Blended Model Metrics:

- Train Accuracy = 74%
- Test Accuracy = 66%
- F1 Score = 65%
  (82%, 54%, 61%, 63%)

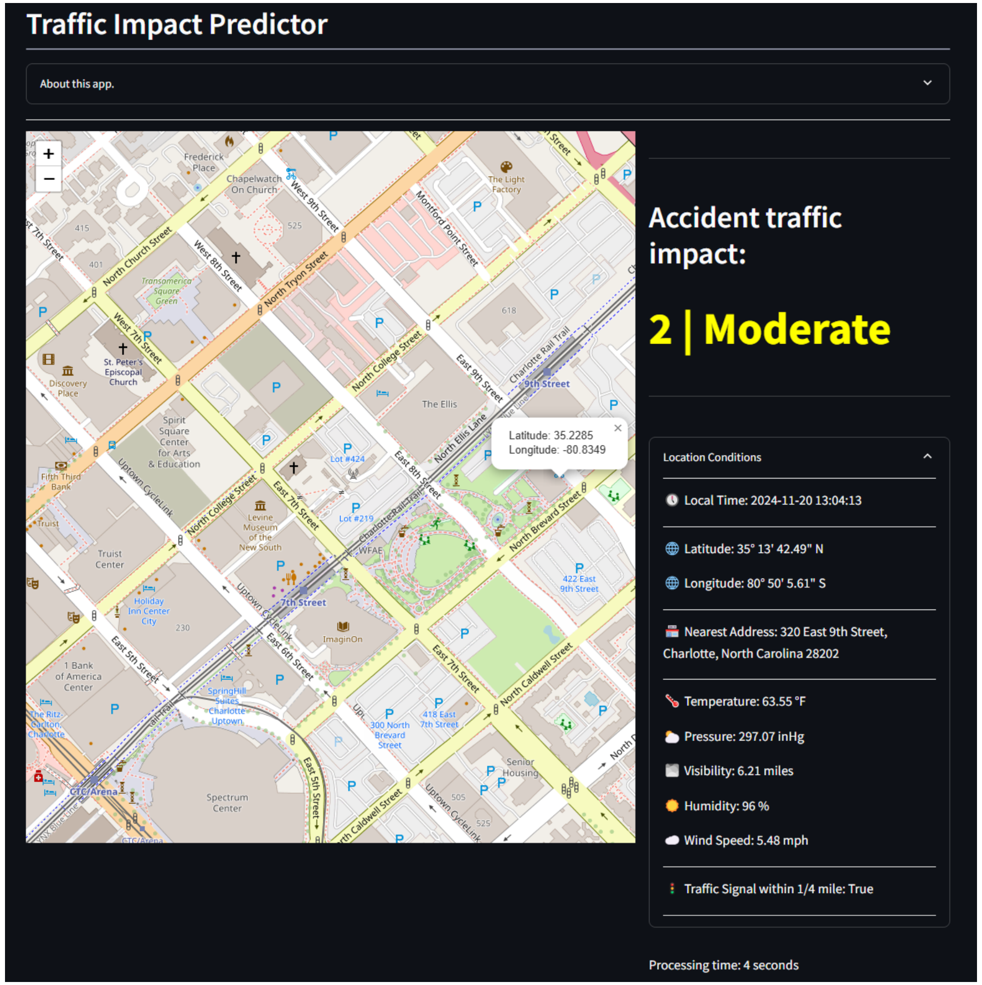
Prediction Dashboard

Purpose: Provide a traffic impact prediction for a user defined accident location

Goal: One Click Prediction

- Minimizes required user inputs
- Doesn't require user to know or look up various feature values
- Improves experience, speeds up delivery of prediction

The App: [Traffic Impact Predictor]



THANK YOU

Q&A

UNIVERSITY OF NORTH CAROLINA
CHARLOTTE