Case Study

Traffic Jam severity Due to Accident

About Data

- Data Indicates Traffic Severity in US post Accident, 1 being lowest sever and 4 is the highest.
- Data has around 49 columns including target variable.
- Data contains numeric columns (Humidity), categorical columns(ex: city), boolean columns(accident side), date columns (ex: start date, end date) and text column(description).

Problem Sizing:

- An unforeseen incident on the road has led to an unexpected obstacle—road closure, resulting in a traffic jam that forces people to endure extended hours of travel, causing significant fuel wastage, affecting work hours in the area, and contributing to noise pollution. Anticipating traffic jams in advance enables the state to gain a better understanding of the demand for resources (such as traffic management) and allows for timely notifications to encourage people to consider alternative routes, thereby minimizing the severity of the situation.

Approach:
Treat problem as multiclass classification. Start with the benchmarking model with mandatory step for modeling. Then Improve model with Feature Engineering.

Solution Steps:

- Check for Class Imbalance
- Missing Value, Constant value, Duplicate Value handling
- Benchmarking model
- EDA and Feature Engineering
- text based feature engineering
- Final Model

Evaluation

Objective Function:

- mlogloss

- merror

Offline Evaluation:

- Balanced classification accuracy
- Confusion metrics
- F1-score

Online Evaluation:

- Reduction in avg delay after accident

Check for Class Imbalance

- Data is Highly Imbalanced with following proportions.

Class	Percentage
1	0.83%
2	67.43%
3	28.43%
4	3.2%

Missing value Imputation

- Remove fields with more than 50% missing value

Field Name	Missing percentage
End_lat	70.45%
Enf_lng	70.45%
Number	64.40%
Wind_Chill(F)	53.17%
Precipitation(in)	57.67%

Constant Value:

- Country and Turning_Loop has constant value for entire dataset and not capable of predicting target variable.
- However, we can remove columns with constant value for most of the records (> 97%) but since data is highly imbalance and tree models are capable of handling imbalance data they are not removed.

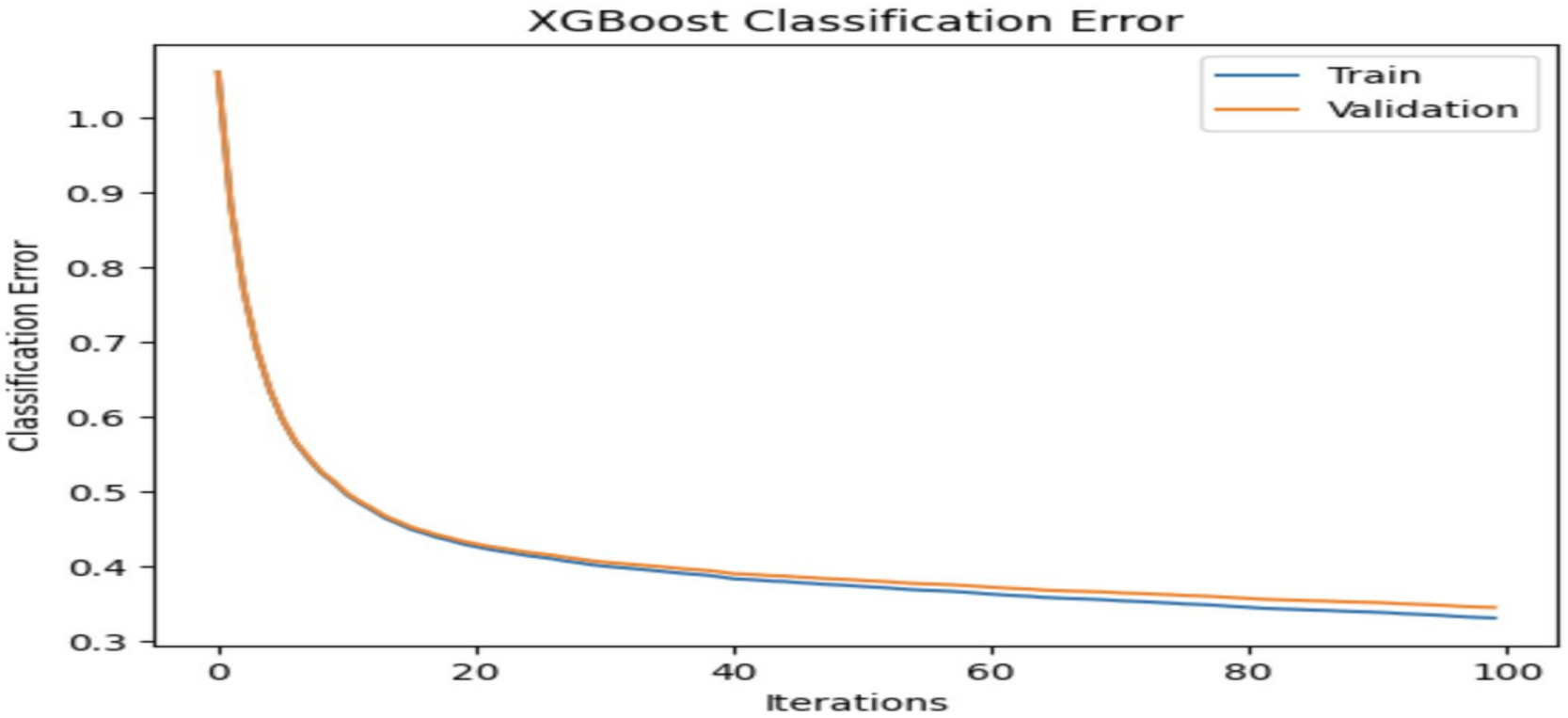
Data Consistency:

- Dates are in different timezones, US operates on different time zone and based on US state open source data, verified that date columns indicates local timezone.

Benchmark Model: Introduction

- Benchmark model can be heuristic or model without any feature engineering and EDA. It’ll help us to evaluate performance enhancement due to feature engineering and feature selection.
- Benchmark model should have better accuracy than random prediction or majority class prediction.

Benchmark Model: Performance



Benchmark Model: Performance

```
.................. balanced_accuracy_score .................
0.8990186534632835
.................. confusion_matrix ..................
[[  6811    380     94      9]
 [ 26344 504508  60205   2246]
 [  4492  22401 219381   3454]
 [   317    558    993  26212]]
.................. classification_report ..................
              precision    recall  f1-score   support

           0       0.18      0.93      0.30      7294
           1       0.96      0.85      0.90    593303
           2       0.78      0.88      0.83    249728
           3       0.82      0.93      0.87     28080

    accuracy                           0.86    878405
   macro avg       0.68      0.90      0.73    878405
weighted avg       0.90      0.86      0.87    878405
```

Feature Engineering: Handle date columns

- Data columns gives temporal attributes like hour, weekday which help understand pattern of severity based on peak hours, office hours and weekdays and weekends. So hour and weekday are parsed from dates and considered as separate columns.

Feature Engineering: Geometric category

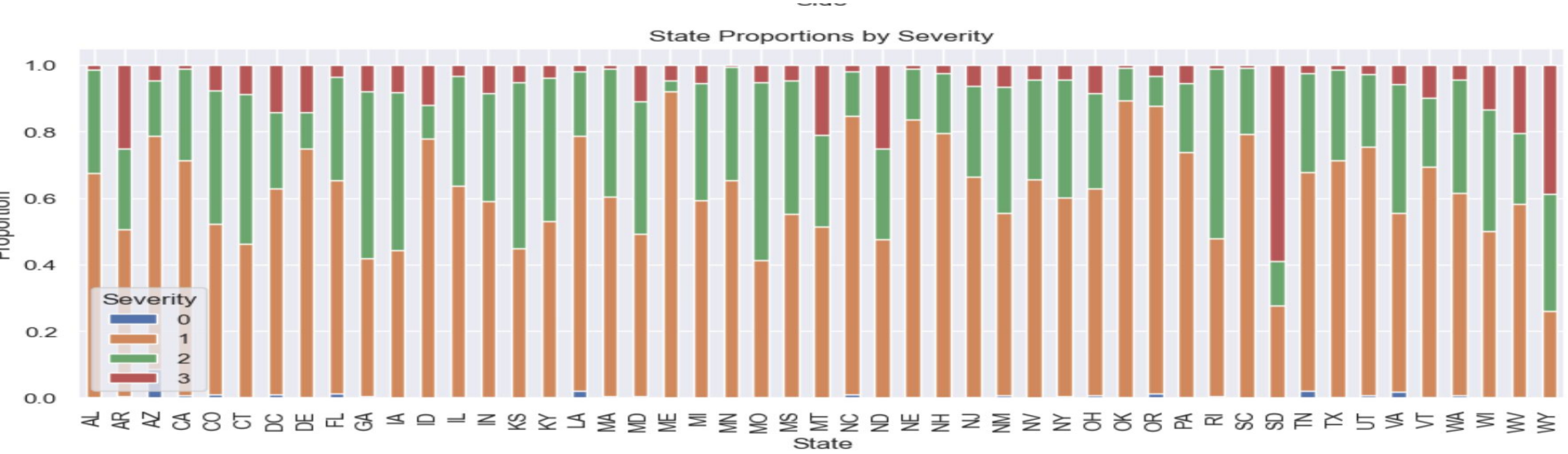
- S2ID10, S2ID11 and S2ID12 are computed from startlat and start_lng. These provides locality categories of different size.

EDA: Understand Accident pattern

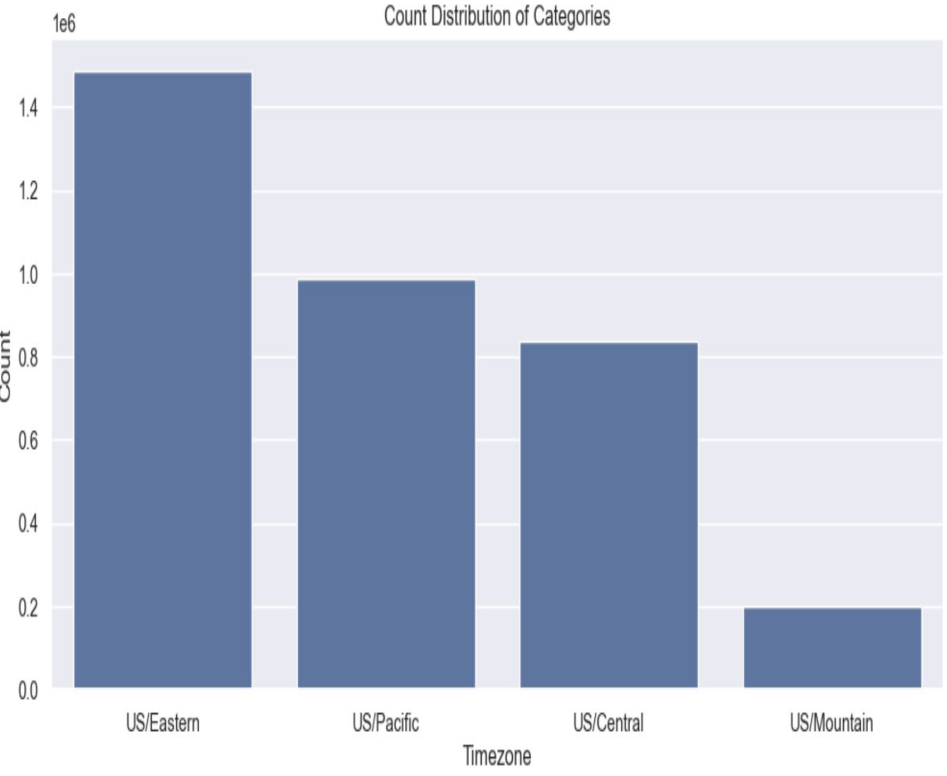
- These dataset is for accidents and density of column value helps with accident states, these helps to identify most accident prune regions, time and days.

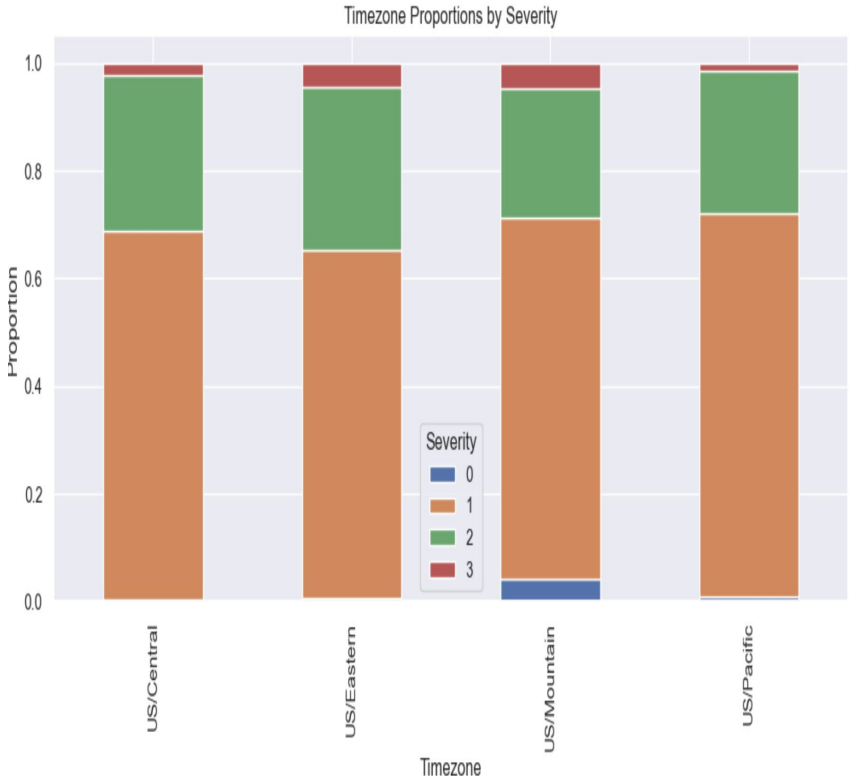
EDA: State wise Accident

- Data Suggest that CA, TX and NC has highest accident when comes to absolute number.
- But Traffic severity proportion is highest for: SD, WY and ND



EDA: TimeZone

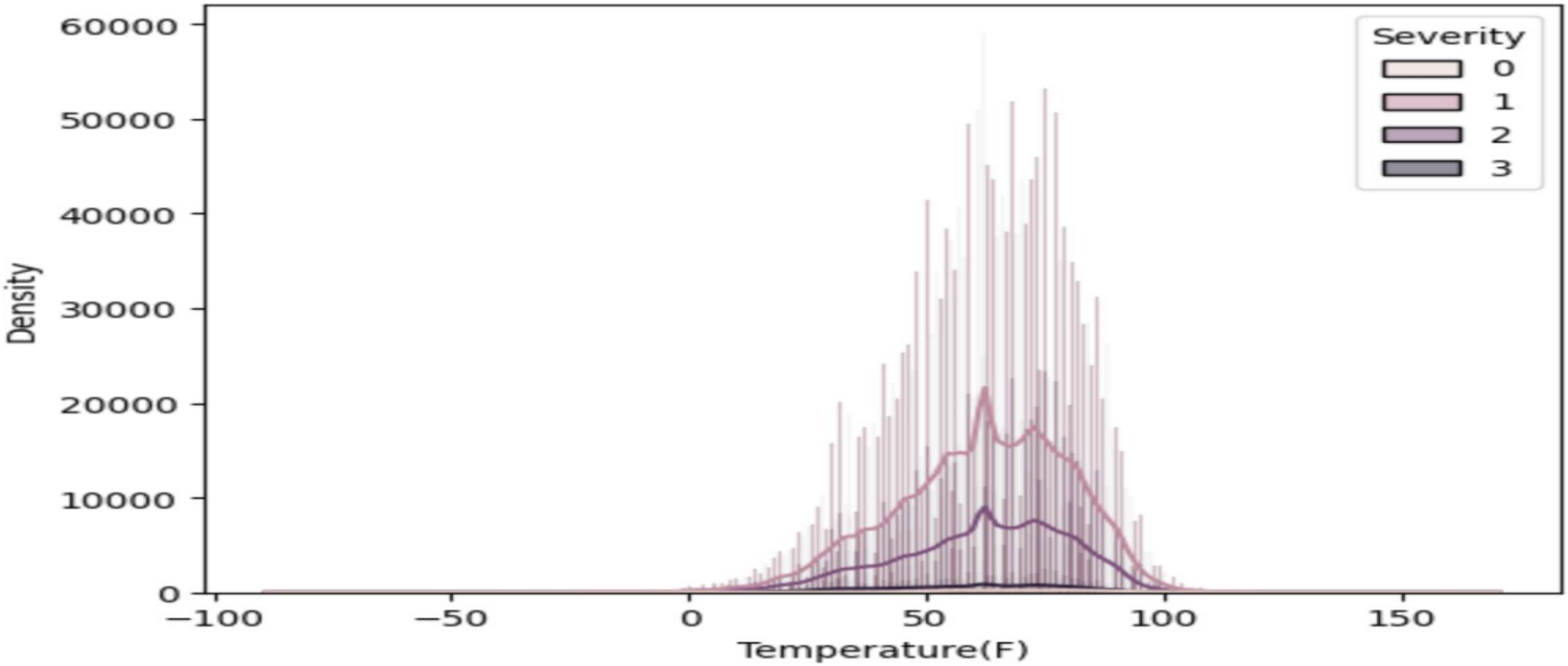




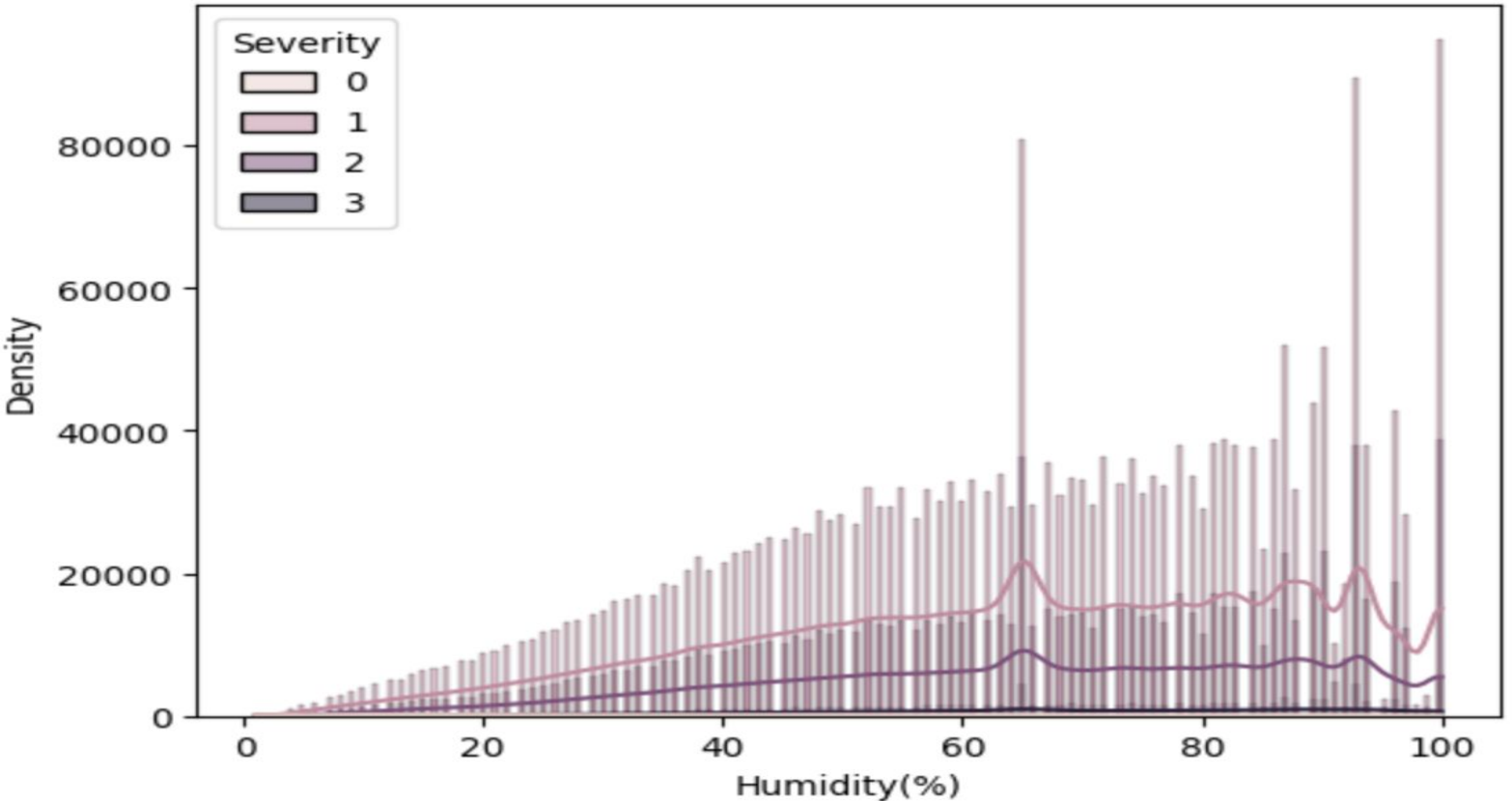
EDA:

- Please follow notebook, for other columns which provide pattern of severity and density distribution.

EDA: Severity Distribution Temperature



EDA: Severity Distribution Humidity



Feature Engineering: Remove correlated feature

- After we add all engineered feature, we expect our features to independent hence correlated features are identified and remove.
- We found hour computed from start date and weather date correlated hence one of them removed.

Feature Engineering: Text feature

- TFIDF vector computed for top 30 words after basics text processing like stopword removal and lemmatization.
- Nltk (as it provided better speed than spacy) used for text cleaning and sklearn for TFIDF vector

Outlier Detection

Column	Number of Outliers
TMC	373200
Distance(mi)	746426
Temperature(F)	26158
Humidity(%)	0
Visibility(mi)	777062
Wind_Speed(mph)	94375

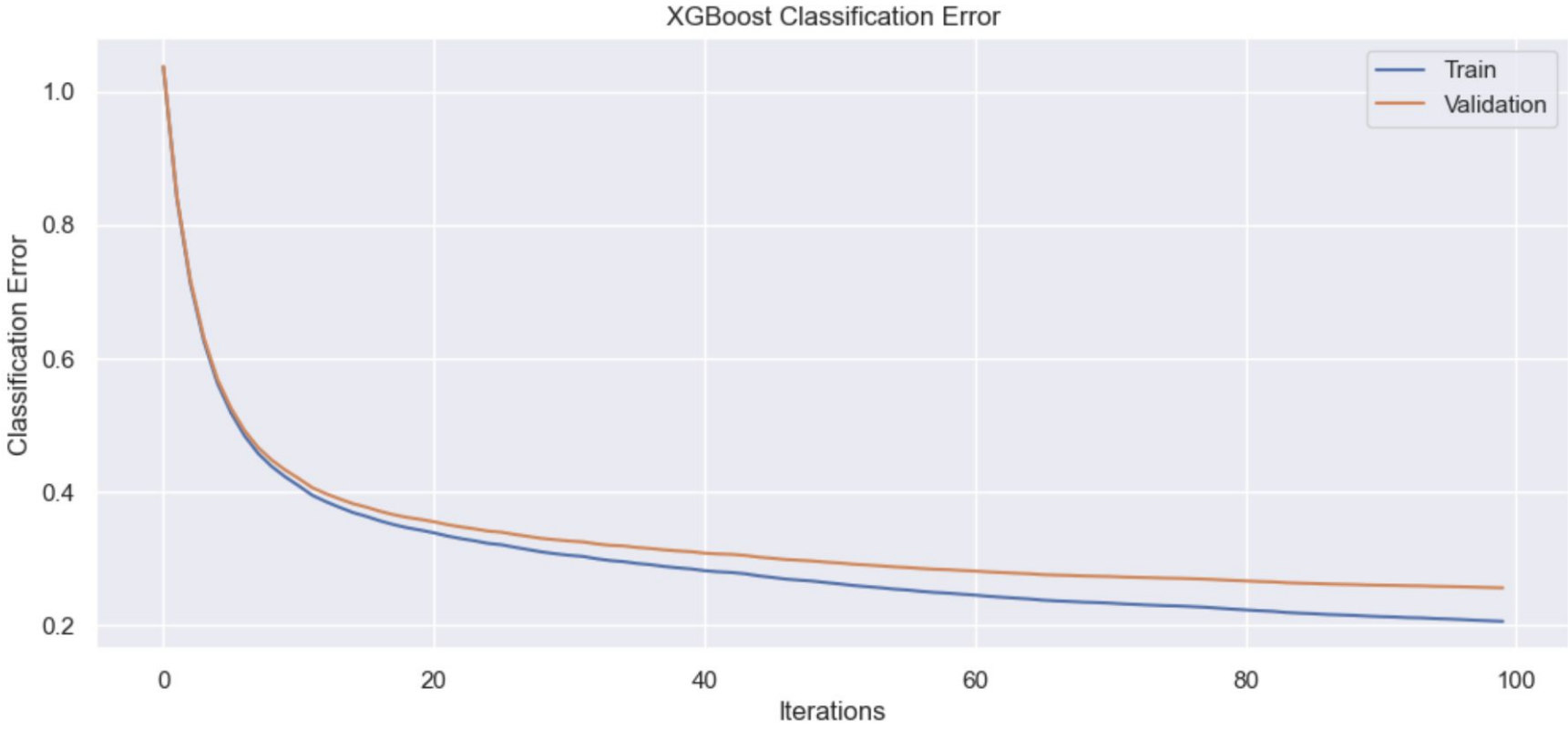
Outlier Treatment

- outlier are Kept as they are more than minority class samples even though tree based model are impacted by outliers but in later version impact of outlier across different class should be analyzed and handled.

Final Model

- As basic model perform and fit well (Does not overfit) on xgboost with considerable parameters. Final model is also Xgboost classifier model.
- Also, Tree base model helps with imbalance data and feature selection and saves considerable feature engineering effort.

Final model evaluation



Final model evaluation

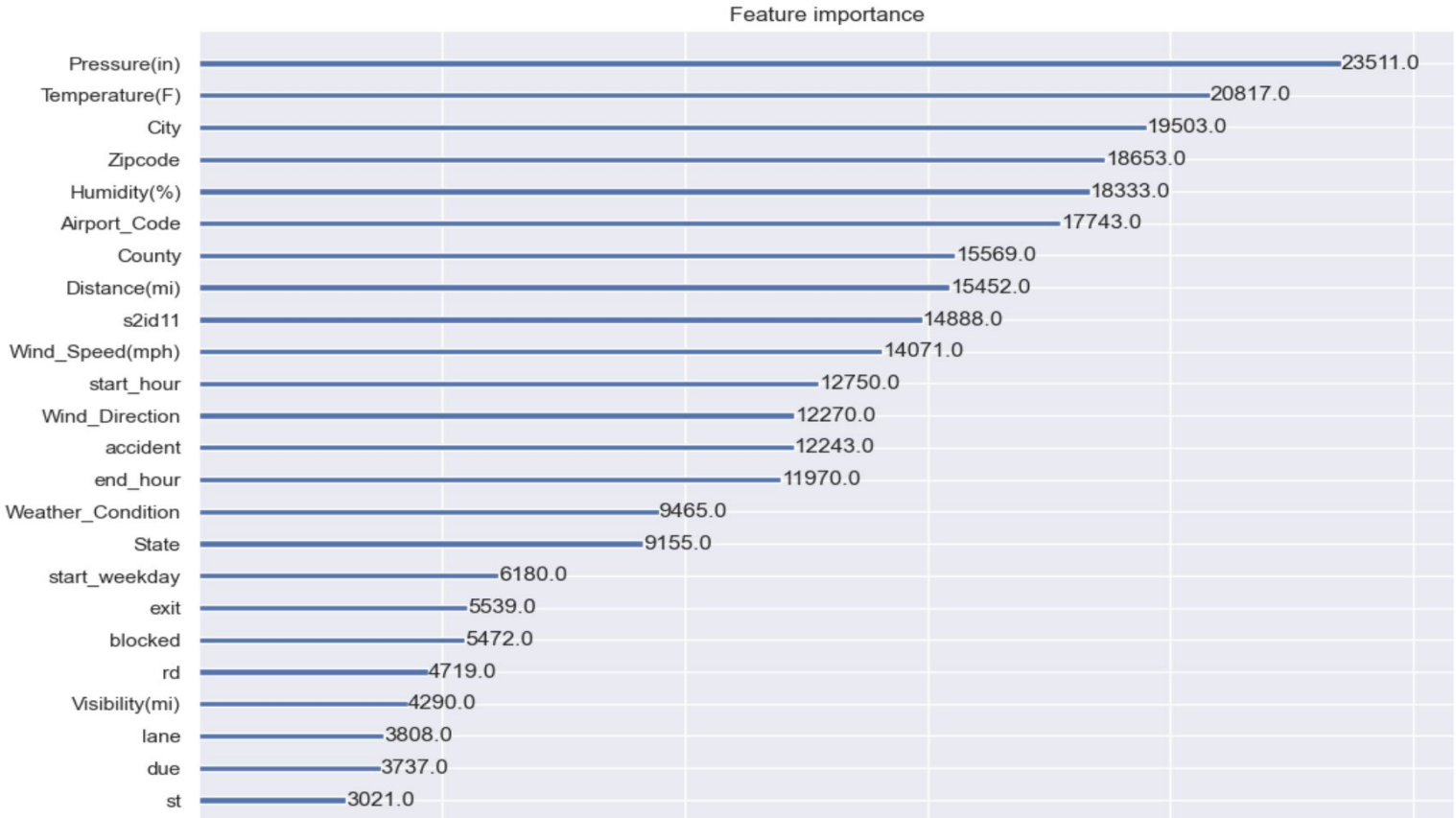
```
.................. balanced_accuracy_score .................
0.9096382455375159
.................. confusion_matrix .................
[[  6601    512    170     11]
 [  6820 532670  53246    567]
 [  1481  21133 225592   1522]
 [   108    416   1374  26182]]
.................. classification_report ...................
              precision    recall  f1-score   support

           0       0.44      0.90      0.59      7294
           1       0.96      0.90      0.93    593303
           2       0.80      0.90      0.85    249728
           3       0.93      0.93      0.93     28080

    accuracy                           0.90    878405
   macro avg       0.78      0.91      0.83    878405
weighted avg       0.91      0.90      0.90    878405
```

[21]

Feature Importance:



Next Steps:

- Deployment Pipeline for online prediction as this is real time prediction use case.
- Although Tree based model performed well on dataset following should be consider for next iteration
- Randomized search CV for hyper parameter tuning
- Use advance embeddings for description as it has good amount of feature importance.
- USe Neural network or more complex model.