Car Accident Severity Prediction

Drive slow

OUR TEAM

Dzhavid Sadreddinov

01

Nikita Borisov

01

Amir Bikineyev

01

Dmitry Dydalin

01

PROBLEM DEFINITION

Objective

Business objectives

Automate the process of obtaining data insights from historical data of car accidents.

ML objectives

Develop a machine learning model to accurately predict the level of the accident severity based on non personal data

Dataset Characteristics

Source: Hugging Face
name of dataset: nateraw/us-accidents
Number of records: ~2.8M

Characteristics of features:

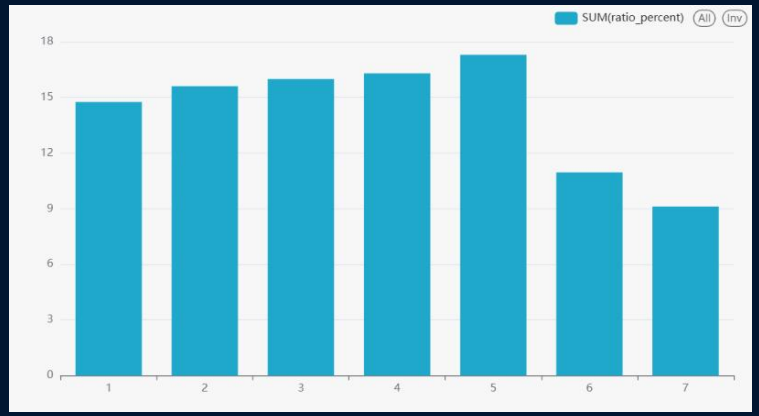
- Number of columns: 22
- Columns containing NaN values: city, state, Weather_Timestamp, temperature, wind_chill, humidity, pressure, visibility, wind_speed, precipitation, weather_condition

Dataset Characteristics

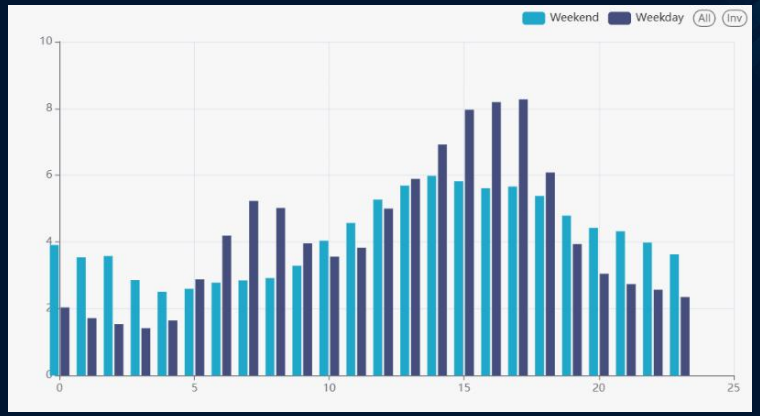
- Numerical columns: Distance_mi, Temperature_f, Wind_chill_f, Humidity_percent, Pressure_in, Visibility_mi, Wind_speed_mph, precipitation_in
- Categorical columns: Side, City, County, Weather_Condition, State
- Geospatial columns: Start_Lat, Start_Lng, End_Lat, End_Lng
- Date column: start_time, end_time, weather_timestamp

EDA

Distribution by Day of the Week

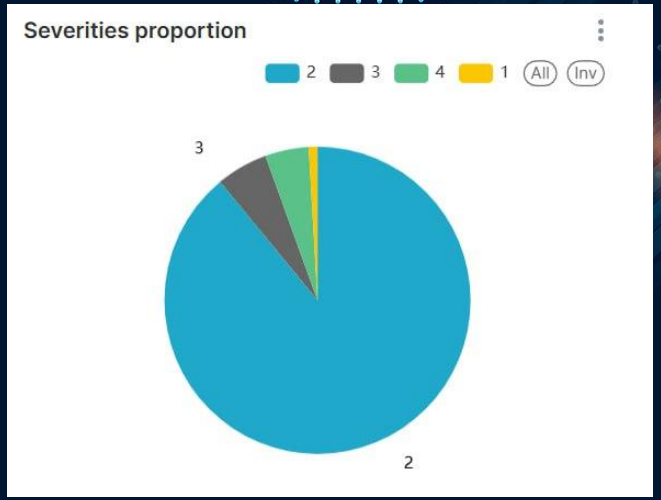


Distributions by hours comparison between working days and weekend

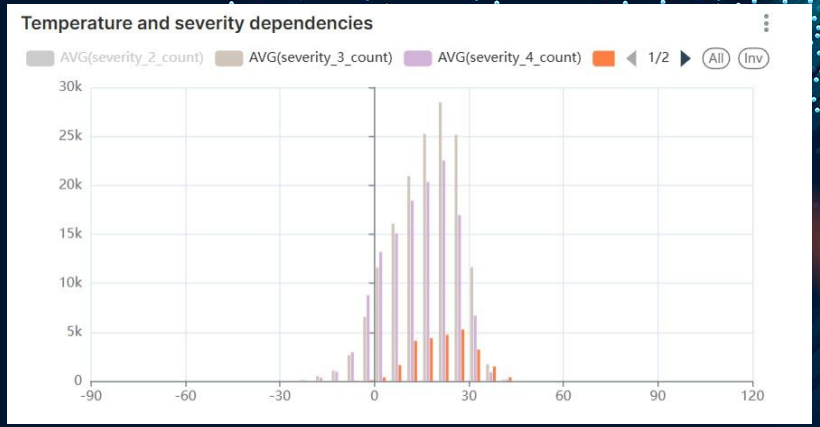


EDA

The results shows that our dataset is highly imbalanced in case of target column. In addition

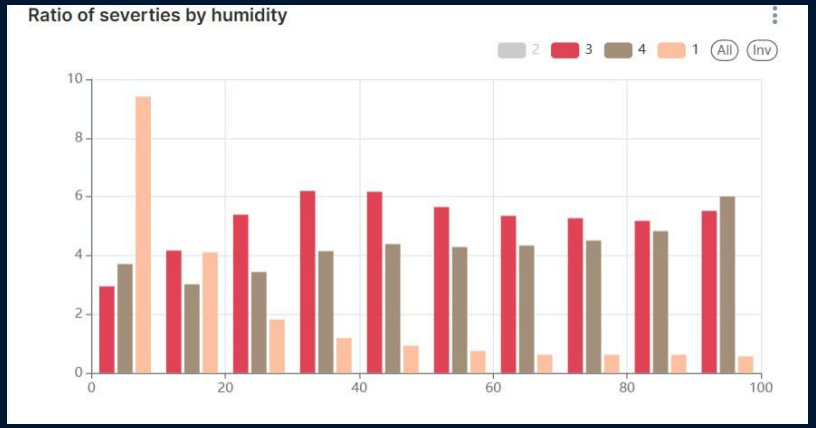


Analysis of results



Analysis shows that severity higher when:

- humidity is high
- temperature is low
- night



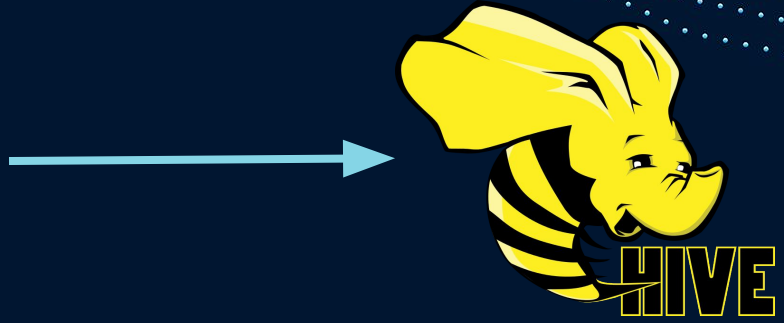
Stage I

Download
.csv dataset

Put the dataset
to PostgreSQL
(batch load)

Create a
compressed
parquet file
on the hdfs

Stage II - Data transfer



We faced an issues with data type matching

It was solved during transition from external to external partitioned and bucketed

Stage II - EDA

All queries were directly linked to the target column ‘severity’. The goal was to find out insights of features that contributes to final results most of all. The visualization was made using Apache Superset ‘charts’ tab

Stage III

1st Model
Logistic Regress

Tuned Params:
regParam
elasticNetParam

Accuracy: 0.891
F1: 0.858

2nd Model
Decision Tree

Tuned Params:
maxDepth
Impurity

Accuracy: 0.898
F1: 0.869

Stage IV

Dashboard: link

Challenges

1. Data Type Incompatibility

2. Resource Constraints on the Hadoop Cluster

3. Superset Limitations

4. Lack of Training Progress Feedback in Spark