US Accident Analysis

Final Term Project

DATS_6450 Cloud Computing

Instructor: Walcelio Melo

by

Group 7

Meghana Gantla

Lokesh Bokkisam

Table of Contents

Scope of the project

- The scope of the project is to build machine-learning models that can predict the severity of an accident based on given features.
- The models will be trained using nearly 2.2 Million rows of data on accidents in different locations and conditions.
- Finally, we want to host a static webpage using S3 to clearly demonstrate the results of our analysis.

Features Implemented

- Preprocessing
- Exploratory Data Analysis
- Feature selection
- Modelling
- HTML, CSS, and JS to access results
- Hosting a static website

Data Source

kaggle

US Accidents (2016 - 2021)

A countrywide traffic Accident Dataset (2016 - 2021)

https://www.kaggle.com/datasets/sobhanmoosavi/us-accidents

About Data

2.8 Million rows

We have a bunch of missing values in numerous columns that need to be handled before performing analysis

47 Columns

- 14 Numeric columns
- 20 Categorical columns
- 13 Binary columns
- 3 Datetime columns

Expected Outcomes

- Understand the distribution of data
- Find out the top locations for accidents
- Find out the prime times for accidents
- Understand features affecting the severity of accidents
- Build machine-learning models for the severity with the reduced feature set

Logical Architecture

Data flow

- The initial dataset has some missing values and a possibility for redundant rows
- We checked for any redundant data using the ‘ID’ feature of the dataset to find and remove all repeated rows
- We analyzed the missing values in each column and decided to deal with them by dropping the columns with more than 30% NA values and then dropping the rest of the rows
- We performed feature selection and reduced the data to 26 columns by removing all binary and some irrelevant columns

Cloud Services

- **Amazon S3:** Amazon S3 (Simple Storage Service) is a cloud-based object storage service provided by AWS that allows you to store and retrieve large amounts of data in a highly scalable and cost-effective way.
- We decided to use an S3 bucket from AWS to host our static website using our HTML webpage and all the related files.

aws Services Search [Alt+S] Global Lokesh_Bokkisam

Upload succeeded
View details below.

Summary

Destination	Succeeded	Failed
s3://term-project	23 files, 4.1 MB (100.00%)	0 files, 0 B (0%)

Files and folders | Configuration

Files and folders (23 Total, 4.1 MB)

Find by name

1 2 3

Name	Folder	Type	Size	Status	Error
image1.jpg	website/	image/jpeg	169.8 KB	Succeeded	-
image10	website/	-	32.5 KB	Succeeded	-
image10.jpg	website/	image/jpeg	32.5 KB	Succeeded	-
image11.jpg	website/	image/jpeg	55.3 KB	Succeeded	-

aws Services Search [Alt+S] Global Lokesh_Bokkisam

Successfully edited static website hosting.

Amazon S3 > Buckets > term-project

term-project Info

Objects | Properties | Permissions | Metrics | Management | Access Points

Bucket overview

AWS Region	Amazon Resource Name (ARN)	Creation date
US East (N. Virginia) us-east-1	arn:aws:s3:::term-project	May 1, 2023, 02:56:30 (UTC-04:00)

Bucket Versioning

Versioning is a means of keeping multiple variants of an object in the same bucket. You can use versioning to preserve, retrieve, and restore every version of every object stored in your Amazon S3 bucket. With versioning, you can easily recover from both unintended user actions and application failures. Learn more

Edit

Bucket Versioning
Disabled

Successfully edited Block Public Access settings for this bucket.

Amazon S3 > Buckets > term-project

term-project Info

Objects | Properties | Permissions | Metrics | Management | Access Points

Permissions overview

Access
Objects can be public

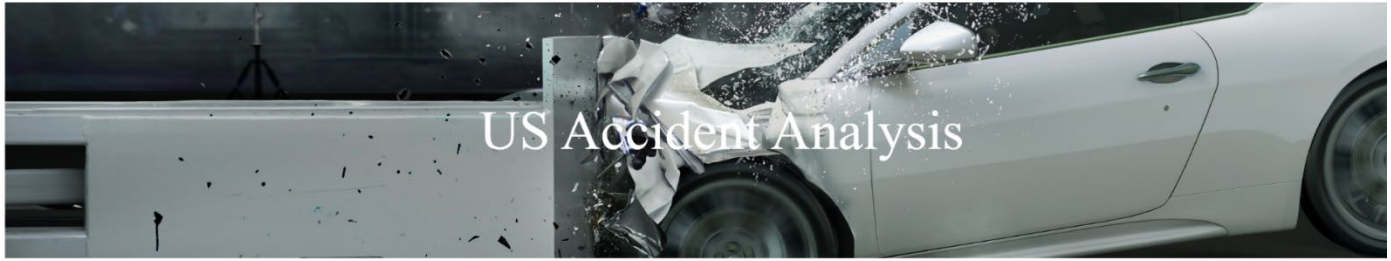
Block public access (bucket settings)

Public access is granted to buckets and objects through access control lists (ACLs), bucket policies, access point policies, or all. In order to ensure that public access to all your S3 buckets and objects is blocked, turn on Block all public access. These settings apply only to this bucket and its access points. AWS recommends that you turn on Block all public access, but before applying any of these settings, ensure that your applications will work correctly without public access. If you require some level of public access to your buckets or objects within, you can customize the individual settings below to suit your specific storage use cases. Learn more

Edit

Outputs

term-project.s3.amazonaws.com/website/index.html



Abstract

This project focuses on understanding the parameters affecting an accident's severity and other constraints like visibility at the time of the incident and the total affected time by the incident. The reason behind selecting this particular dataset is mainly its usability and detailed columns. Many thousands of accidents happen each day across the world and they vary in severity and reason. In this analysis, we understand the constraints that affect these incidents and also find out the hotspots.

Description of Dataset

The dataset is from Kaggle and has over 2.8M rows with 47 columns. In the data, we have 14 numeric columns, 20 categorical columns, and 13 binary columns. These columns also include 3 DateTime columns that are under object datatype. The description of columns after preprocessing is as follows,

- **Index:** The index of the row.
- **Severity:** A number from 1 to 4 that represents the severity of the accident
- **Start_Lat:** The latitude of the starting point of the accident.

Models and Accuracy scores

We have tried out 4 different models with Severity as our target column. We had Start_Lat, Start_Lng, End_Lat, End_Lng, Distance, Temperature, Wind_Chill, Humidity, Pressure, Visibility, Wind_Speed, Precipitation, Start_Time, End_Time, Street, Side, City, County, State, Timezone, Airport_Code, Weather_Timestamp, Wind_Direction, Weather_Condition, and Sunrise_Sunset in our training data. All the models have given more than 90% accuracy score which is really good. With the results from analysis, we will be able to predict the severity of an accident given rest of the conditions with really good accuracy.

	Models	**Accuracy score**
1.	DecisionTreeClassifier	0.92
2.	LogisticRegression	0.93
3.	KNeighborsClassifier	0.94
4.	RandomForestClassifier	0.95

Conclusion

- California and Florida are the states with the highest number of accidents in this dataset. However, we can also observe that there isn’t enough data from a few states.
- Severity of an accident can be predicted with over 90% accuracy with the help of the remaining information in the dataset.

Thank you!