Building an Accident Risk Prediction Model Using Machine Learning Algorithms on Heterogeneous Sparse Data

Anirudh Srinivasan
Ira A.Fulton Schools of Engineering
Arizona State University
Tempe, USA
anirudhs@asu.edu

Manohar Akula
Ira A.Fulton Schools of Engineering
Arizona State University
Tempe, USA
makula@asu.edu

Pranav Murali
Ira A.Fulton Schools of Engineering
Arizona State University
Tempe, USA
pmural5@asu.edu

Srikanth Sampath
Ira A. Fulton Schools of Engineering
Arizona State University
Tempe, USA
ssampa28@asu.edu

Key Words—Risk Prediction, Logistic Regression, Decision Tree, Random Forest, KNN, MLP Classifier-Neural Network

I. Problem Formulation

A. Problem Statement

Road traffic crashes are the leading cause of deaths today in the United States. CDC reports an alarming statistic that the United States witnesses nearly 100 deaths each day. It is therefore the need of the hour to leverage research and analysis of real-time traffic and accident data to predict the risk of accidents. Accident risk prediction can help in improving the public safety infrastructure by predicting and warning the public in advance. In this project, the severity of accident risk for a given set of conditions is predicted. Various features like weather, traffic volume, road conditions, time of the day, description of the previous accidents are analyzed from the dataset.

Given the road and environment conditions, our expected results would help predict the severity of accidents in the United States. We are going to be presenting our project to the class with an illustration of all the machine learning model predictions through a collection of videos and interactive plots. We would also conclude by showing which model and algorithm works best in our project.

B. Workflow Model

The workflow model adopted for this project is split into four steps. The first step involves reading the input data in the required format. In the second step, we then move on to check if there are any discrepancies in the data and identify what processes must be done to correct them. We then clean the data by removing the outliers, imputing the missing values and transforming the data. The third step involves exploring the cleaned data set to find any initial data patterns and identify useful insights by building visualizations.

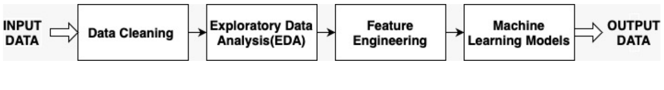

Fig. 1.Workflow mode

Using these insights, we move on to the next step, feature engineering. Here the parameters that are most ideal for the model are determined and the necessary preparation to feed the data into the model is carried out. In the final step, we split the data into train and test datasets. We build the machine learning model using the train dataset on the features we have selected and finally predict the output for the test dataset.

C. Key Concepts

This project aims to build an accurate model to predict the accident risk in the given conditions. In this section, we outline the various machine learning models that were considered for this purpose.

i. Logistic Regression

The Logistic regression class is imported from the sci-kit library. Multinomial logistic regression is a classification method that generalizes the algorithm to multi-class problems. It is a model that is used to predict the probabilities of the different possible outcomes of a categorically distributed dependent variable, given a set of independent variables.

ii. K-Nearest Neighbors Model

K-Nearest Neighbor (KNN) is a machine learning algorithm based on the supervised learning technique. It encompasses data points into the available categories based on similarity. It is a non-parametric algorithm, meaning it does not make any assumption on the underlying data. It does not learn from the training set immediately instead it stores the dataset and at the time of classification, it performs an action on the dataset.

iii. Decision Tree

We import the Decision Tree class from Sci-kit library, this model predicts the value of a target variable based on several input variables(features). A tree is built by splitting the source set, constituting the root node of the tree, into subsets. The splitting is done based on a set of rules by classification features. This process is repeated on each derived subset in a recursive manner until the subset at a node has all the same values of the target variable, or when splitting no longer adds value to the predictions.

iv. Random Forest

Random Forest is a machine learning technique that is used to solve regression and classification type problems. It utilizes ensemble learning combining classifiers to provide solutions for many complex problems. A random forest consists of many decision trees. The algorithm establishes the outcome based on the predictions of the decision trees. It predicts by taking the average or mean of the output from various trees.

v. MLP Classifier - Neural Network

It stands for Multi-layer Perceptron classifier which in the name itself connects to a Neural Network. MLP is a type of artificial neural network (ANN). Simplest MLP consists of at least three layers of nodes: an input layer, a hidden layer and an output layer. Unlike other classification algorithms such as Support Vectors or Naive Bayes Classifier, MLPClassifier relies on an underlying Neural Network to perform the task of classification.

II. Specific References

A. Key literature

1) Sobhan Moosavi, Mohammad Hossein Samavatian, Srinivasan Parthasarathy, Rajiv Ramnath. 2019. A Countrywide Traffic Accident Dataset. arXiv:1906.05409v1 [cs.DB]. This paper compiles a large scale publicly available dataset of accident information through a comprehensive process of data collection, integration and augmentation.

2) Moosavi, Sobhan, Mohammad Hossein Samavatian, Srinivasan Parthasarathy, Radu Teodorescu, and Rajiv Ramnath. "Accident Risk Prediction based on Heterogeneous Sparse Data: New Dataset and Insights."arXiv:1909.09638v1 [cs.LG]. This paper then builds different models using the features identified to predict the accident risk for given conditions.

B. Datasets

The dataset used is taken from the a publicly available open source dataset. It consists of a country wide US accident dataset which covers 49 states of the USA. This dataset was compiled using multiple APIs which provide streaming traffic incident data. These APIs broadcast traffic data captured by a variety of entities, such as the US and state departments of transportation, law enforcement agencies, traffic cameras, and traffic sensors within the road-networks.

Dataset Link:

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

C. Code Implementation

The code we have taken reference from is the model implemented by Naga Janaki Dwadasi. We have executed the code successfully and confirmed that it produces acceptable results. From this code, we have built different machine learning models mentioned to compare accuracies and select the most accurate one.

Code Link:

- https://github.com/NagaJanakiDwadasi/US-Accident-Risk-Prediction

Working Code:

- https://drive.google.com/file/d/18tzM4qIepXIkhnNEFw4cMZQVJrx37bEp/view?usp=drivesdk
- https://drive.google.com/file/d/1uB6W_yzQt-2bf6THrJyBvffOY2iDqUQ3/view?usp=sharing
- https://drive.google.com/file/d/1lyDRKUR1kt8vJg0WBm6ZuLwFtPJ92KiI/view?usp=sharing

III. Procedure

A. Project Overview

The reference paper focusses on the creation of the dataset using comprehensive process of data collection, integration and augmentation. In addition to this, we use this dataset to perform exploratory data analysis to identify data patterns and gain significant insights. We then perform feature engineering to select the best features relevant for our model prediction. We further enhanced the idea in the paper, by implementing additional machine learning models to predict the the probability of accident risk given the parameters. We implement models like Logistic Regression, Decision Tree, K-Nearest Neighbors, Random Forest and MLP Classifier Neural Network and then their accuracies were compared to select the best accident risk prediction model.

B. Software Requirements

- Computing platform: If running on local system, 12GB RAM laptop, MacOS/WindowsOS, 6GB Graphics, Python 3 installation, Anaconda - Jupyter Notebook (OR)
- Google Colab: 10GB Google Drive space for mounting

C. Data Cleaning

The raw data we have in our dataset must be cleaned in order to observe significant patterns to get insights from the data. We detect and correct corrupt or inaccurate records by replacing or deleting them. We handle null values by imputing appropriate values (impute the mean in case of numerical features and most common value in case of nominal features), so that we do not have data loss by just deleting null entries.

D. Exploratory Data Analysis and Illustrations

Exploratory data analysis was then carried out to discover patterns, test hypothesis and identify relations between different attributes of our dataset. This was achieved by using summary statistics and various graphical representations.

The following are some examples of insights gained through our exploratory analysis:

i. From the map below, it is evident that the number of accidents in the country is relatively higher along the eastern and western coast

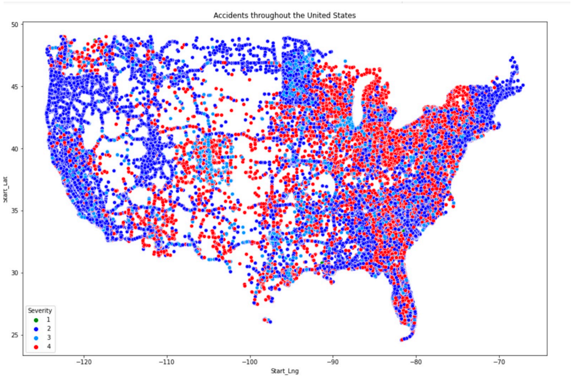

Fig. 2. Region wise Accident Distribution

ii. From the below plot, we see that 7 of the top 10 states are coastal states which corroborates our previous finding. We can also see that California experiences significantly higher number of accidents in comparison to others.

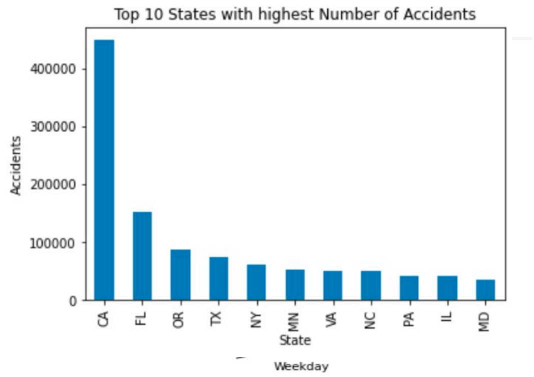

Fig. 3. Top 10 states with most number of accidents

iii. A closer look into the accident distribution by severity indicates that there is there significantly more accidents of severity 2

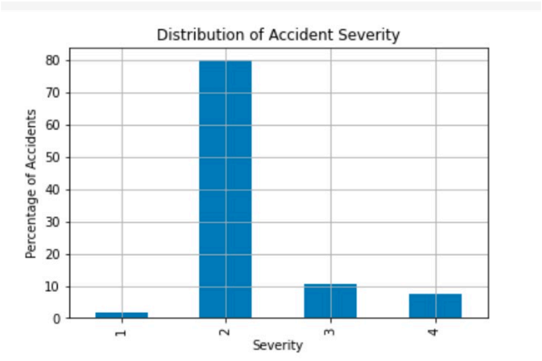

Fig. 4. Accident Distribution by Severity

iv. Taking a look at the distribution of the accidents over the week, we find that the number of accidents is always high over the weekdays in comparison to the weekends

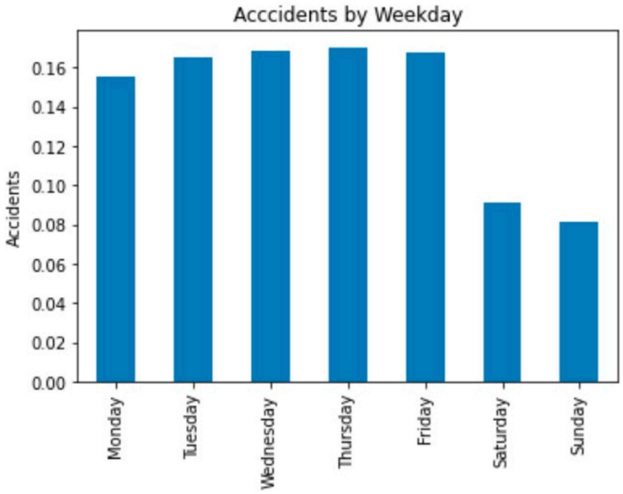

Fig. 5. Accident Distribution over the Week

E. *Feature Engineering and Modeling*

It is the process of using domain knowledge to extract features from raw data. We plot the correlation matrix to find the relationship between the variables. Next we split the dataset into train and test. Further, we train the model, test its accuracy and make predictions.

IV. Future Progress

Between progress check-1 and progress check-2, we are going to be building the machine learning models for enhancing the accuracy further, working on hyperparameter tuning. We are going to be involved further in data validation, data visualization as well. Our milestones would be to enhance the built models for better accuracy, come up with innovative and exciting data visualization plots to present to the class and make a standardized report which would explain our project perfectly to both a beginner and a professional on the same level.

V. Individual Responsibilities

Anirudh and Pranav would be responsible for hyperparameter tuning and enhancing the accuracy of KNN, Logistic Regression and Decision Tree models.

Srikanth and Manohar would be responsible for hyperparameter tuning and enhancing the accuracy of MLPClassifier, Random Forest models.

All of us would be equally contributing towards the final project report upon consultation with each other so that each person puts the appropriate effort towards the subject.

every team member is required to complete this form, which will be reviewed and signed by other team members								
by signing this form, you agree that the student 1) has contributed sufficiently to the team work, and 2) what is claimed by the student is accurate.								
attach this form at the end of your work report for each and every assignment. This form does not contribute to the page count.								
Assignment name	Progress Check 1							
Team #	6							
Student name: Anirudh Srinivasan	worked on literature	worked on implementation (data, platform, test run, debug, compatibility…)	generated results (run results, result data processing, presenting results	wrote report (Intro, method, result, discussions, …)	other significant contributions	peer approval 1	peer approval 2	peer approval 3
specific & detailed evidence is required to support claims of contributions (make reference to equation #, figure #, paragraphs, sections, etc…)	Studied and analyzed "A Countrywide Traffic Accident Dataset" paper. Utilized the cleaned dataset to perform univariate and bivariate analysis to find data patterns.	Setup the environment with the necessary libraries. Debugging for Python's version compatibility for multiple notebooks. Test run "logistic regression" notebook successfully.	Presented output after training the logistic model. Measured accuracy and predicted accident risk	Documented "Problem Statement" and "Workflow model" section	Recorded video for logistic model	Manohar	Pranav	Srikanth
Student name: Manohar Akula	worked on literature	worked on implementation (data, platform, test run, debug, compatibility…)	generated results (run results, result data processing, presenting results	wrote report (Intro, method, result, discussions, …)	other significant contributions	peer approval 1	peer approval 2	peer approval 3
specific & detailed evidence is required to support claims of contributions (make reference to equation #, figure #, paragraphs, sections, etc…)	Studied and analyzed "A Countrywide Traffic Accident Dataset" paper. Obtained the dataset and performed necessary data cleaning tasks	Setup the environment with the necessary libraries. Test run the "Preprocessing" notebook successfully.	Observed the correlation matrix and performed feature engineering to train the models.	Documented "Key concepts" section	Took screen prints of illustrations	Anirudh	Pranav	Srikanth
Student name: Pranav Murali	worked on literature	worked on implementation (data, platform, test run, debug, compatibility…)	generated results (run results, result data processing, presenting results	wrote report (Intro, method, result, discussions, …)	other significant contributions	peer approval 1	peer approval 2	peer approval 3
specific & detailed evidence is required to support claims of contributions (make reference to equation #, figure #, paragraphs, sections, etc…)	Studied and analyzed "Accident Risk Prediction based on Heterogenous Sparse Data". Handled data transformation and anomalies in the dataset.	Setup the environment with the necessary libraries. Test run the "Decision Tree" and "Random Forest" notebook successfully.	Displayed output after training the Decision Tree and Random Forest models. Measured accuracy and predicted accident risk	Documented "Specific references" and " Procedure" section.	Recorded video for Decision Tree and Random Forest models	Anirudh	Manohar	Srikanth
Student name: Srikanth Sampath	worked on literature	worked on implementation (data, platform, test run, debug, compatibility…)	generated results (run results, result data processing, presenting results	wrote report (Intro, method, result, discussions, …)	other significant contributions	peer approval 1	peer approval 2	peer approval 3
specific & detailed evidence is required to support claims of contributions (make reference to equation #, figure #, paragraphs, sections, etc…)	Studied and analyzed "Accident Risk Prediction based on Heterogenous Sparse Data" paper. Explored data patterns and visualized them to identify relationships.	Setup the environment with the necessary libraries. Test and run the "MLP Classifier" and "KNN" notebook successfully.	Visualized output after training the MLP and KNN model. Measured accuracy and predicted accident risk	Documented " Procedure" and "Future progress " section.	Recorded video for MLP model	Anirudh	Manohar	Pranav