Heat-map visualization of traffic accidents in Los Angeles using predictive modeling.

The motivation for this project stems from Los Angeles' reputation as a city with a great deal of traffic and accidents. According to Go Safe Lab's research, Los Angeles is the U.S. state with the third highest counts of traffic accidents.[1] Various navigation applications, such as Google Maps and Waze, have built-in traffic monitoring tools that might indicate to the user the potential for accidents to occur. However, our focus is to build out a predictive model that could instead directly warn drivers of potential hot-spots for accidents.

Along with the reputation for being a traffic-heavy city with congested roads most times of the day and all through the week, Los Angeles has had an increasing number of traffic accidents[2] in the past five years and this is bound to grow with the car sector booming into the electric vehicle era. There are apps mentioned above that monitor traffic and guide vehicle's on the road through less congested areas, however, there are no apps that predictively tell a driver which streets are more prone to accidents of any kind (pedestrian, bicycle, vehicle collisions, etc.). Thus, we want to build a simple and functional app that enables the LA citizen to drive defensively and with a data-driven approach that enables them to avoid unforeseen traffic incidents while commuting.

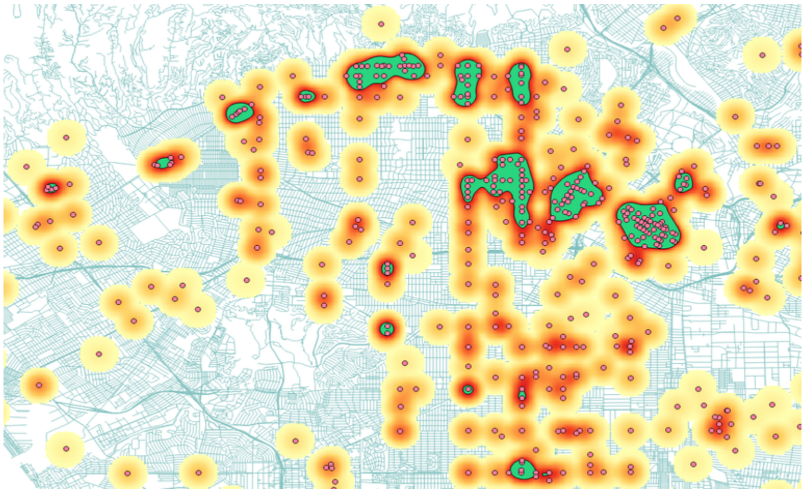

Figure 1: Example Traffic Heatmap

Our first step is to choose a dataset that we can use to build out our model. The dataset we have initially chosen is LA traffic accidents from Los Angeles Open Data.[3] The data corresponds to roughly 500,000 accidents recorded from Jan 2010 until Jan 31 2022. There are a variety of attributes corresponding to each accident. Secondly, we need to perform exploratory data analysis on variables of interest, identifying missing values, outliers and data distributions. For the location coordinates, missing values will be filled by using the geopy library to convert address names to coordinates. Next, pertinent variables are chosen based on domain knowledge and common knowledge. For example, attributes such as "ID number" and "gender" will likely not be significant in relation to accidents, whereas the coordinates and date/time likely are.

[1] https://gosafelabs.com/research/review-of-2019-accident-data

[2] https://www.latimes.com/california/story/2022-01-09/traffic-deaths-vision-zero-garcetti#:~:text=According%20to%20Los%20Angeles%20Police,the%20same%20period%20in%202020.

[3] https://data.lacity.org/Public-Safety/Traffic-Accidents-by-date/2mzm-av8

Additionally, since we are a 3-person team, we will need a secondary dataset to incorporate into our application. In order to do so we have decided to implement a separate heatmap for traffic congestion through data provided on LA Mayor's website.[4] Through these features, drivers will be able to make effective real-time decisions and we aim to promote safe and defensive driving through the use of our app. As mentioned above, Los Angeles has a reputation for accidents and we believe that our app has a real social impact on the well-being of the citizens of Los Angeles.

App Features

1. Unsupervised Learning Approach for Accident Heatmaps

Since each observation corresponds to an accident, there is no data corresponding to the alternative, no accident. Therefore in the context of our application we do not have labeled data to predict accident/not, meaning that we should focus on an *unsupervised* learning procedure. Our goal is to segment data into low/medium/high risk groups or clusters so we must identify which combinations of variables result in the most efficient and effective clustering.

The main challenge we will face regarding this approach will be evaluating the quality of our clusters. In particular, we will have to manually interpret what attributes and their values make a cluster low/medium/high risk. To evaluate the quality of clusters, we aim to use a decision tree tool to understand how the clustering process took place, removing the black-box situation that existed during the unsupervised process.[5] To implement a decision tree, we will follow the assumption that each cluster is a class and assign the observations belonging to a cluster with a corresponding class label. This 'labeled' data can now be fed into the decision tree where we can gather insights on what combination of features leads to low/medium/high risk accident events. However, we are still researching techniques to estimate the accuracy of an unsupervised model.

2. Traffic Congestion Heatmap

While accident data provides a lot of insight to the driver, we believe that integrating it with normal traffic data will benefit drivers even more. Using Tableau, we will be able to create geographical heat maps that provide the user with a feature to quickly switch to see how traffic is in the area. Moreover, we can potentially find a causal relationship between traffic congestion and accidents which can also be implemented into our unsupervised learning approach to create clusters for each street. Like, in the previous feature, the heatmaps will indicate red for high traffic areas, yellow for average traffic and green for

[4] https://data.lacity.org/Transportation/LADOT-Traffic-Counts-Summary/94wu-3ps3
[5] https://towardsdatascience.com/interpretable-clustering-39b120f95a45

low traffic areas. We strongly recommend that these two features be used in tandem to create a safe driving environment for everyone in the city.

Software & Libraries

1. Dataset Management
    a. Firebase: This cloud-based platform will be a great tool for storing the CSV files and will be accessed during the data analysis and front-end stages.
2. Data Analysis (Preprocessing and ML)
    a. Python: Python is our back-end tool since our preprocessing, clustering models and even some parts of front-end design will be executed.
    b. Libraries:
        i. Scikit-Learn: Clustering models, decision trees and evaluation metrics
        ii. Pandas: Data manipulation for preprocessing
        iii. NumPy: Data manipulation for ML tasks
        iv. Joblib/Pickle: Saving the ML models for later usage
        v. GeoPy: Geospatial package for translating address into coordinates
3. Visualization
    a. Tableau: Tableau will be used along with other visualization packages in Python to create effective geographical heat map visualizations that will be implemented into the application.
    b. Folium: A Python visualization package for translating coordinates into map representation
    c. Matplotlib: A Python visualization package to visualize data characteristics
4. Front-End
    a. Streamlit: Front-end Python package for designing and deploying web-based UI
5. Documentation
    a. GitHub: We will use GitHub for version control of all files.

Teammates & Responsibilities

Member	**Background/Skills**	**Responsibilities**
Sofian Ghazali	Undergrad Major: B.S. in Electrical Engineering Skills: Python, SQL, ML models, Tableau, HTML, CSS, JS	- Application of Machine Learning Model - Preprocessing of geospatial data - Optimization of ML Model
Kyle Brooks	Undergrad Major: B.S. in Chemistry Skills: Python, ML models, SQL, EDA, deep learning	- EDA on accident dataset - Assist in model development - Research streamlit documentation and create UI
Abhinav Rao	Undergrad Major: B.S. in Chemical Engineering Skills: Python, SQL, ML Models, Tableau, EDA, GitHub, Pandas, NumPy, Matplotlib, Seaborn, Plotly	- Data Preprocessing - Exploratory Data Analysis - Application of Machine Learning Model - Deployment of app via streamlit

Milestones

Member	**Task**	**Timeline**
Sofian Ghazali	- Data Extraction & Storage	- Feb 10th - Feb 17th
Kyle Brooks	- Data Preprocessing	- Feb 17th - 28th
Abhinav Rao & Kyle Brooks	- Exploratory Data Analysis	- Feb 28th - March 7th
Sofian Ghazali, Abhinav Rao & Kyle Brooks	- Application of Unsupervised Learning Algorithm	- March 7th - March 21st
Sofian Ghazali, Abhinav Rao & Kyle Brooks	- Optimization of Model	- March 21st - April 4th
Kyle Brooks, Sofian Ghazali & Abhinav Rao	- UI Design & App Deployment	- April 4th - April 20th
Kyle Brooks, Sofian Ghazali & Abhinav Rao	- Documentation	- Ongoing Process