NAVIGATING CHICAGO: Analysis of CTA Transit Route

Juveriya Fatima
A20528182
jfatima1@hawk.iit.edu

Vidheesha Patil
A20517203
vpatil21@hawk.iit.edu

PROJECT PROPOSAL

ABSTRACT:

The Chicago Transit Authority operates on multiple routes connecting a wide area of Chicago city and Chicagoland area.This research will study data from various sources and will include factors of passenger footfalls, ticket sales, train and bus schedules, crime by using data analytic techniques such as regression analysis, time series analysis and data visualization. The research aims to identify efficient and secure transit routes for CTA passengers, offering recommendations for a safer, quicker, and more comfortable commute.

RESEARCH GOAL:

The goal of this project is to find the optimum transit routes and timings for Chicago Transit Authority (CTA) passengers by analyzing data from different sources such as traffic, weather, crime and other datasets. To determine the most efficient and secure transit routes for CTA users, our research involves analyzing multiple datasets through techniques such as regression analysis, time series analysis, and data visualization. Furthermore, the project offers suggestions for enhancement of services derived from the results of data analysis. Commuters of CTA and the chicagoland community will benefit in terms of ease of access to safe and efficient public transportation service.

RESEARCH QUESTIONS:

1. What factors (Location, weather condition, facilities, time) that affect the footfall in public transport in Chicago city?
2. Which transit routes have most passengers getting on board?
3. What time-slots are the traffic maximum and minimum in?
4. Comparing the number of trips on different routes between weekdays and weekends.

5. The influence of holidays and weekends on the utilization of the Chicago Transit Authority.
6. The impact of crime statistics in a specific area and the route number on the frequency of trips conducted.
7. Recommendations to CTA on routes with high passenger demand and insufficient transit capacity, considering crime statistics and proposing solutions for effective alleviation.
8. Diagrams based on location & density for the busiest stations.
9. Examining data specific to gender for each of these routes.
10. Could an interactive report (visualization and recommendation) help the authority assess the current condition of Chicago transit use?

PROPOSED METHODOLOGY:

In order to answer the above questions, below is the proposed methodology. We will be following the 3 steps of Data Collection, Data preparation and Analysis on the data.

DATA COLLECTION:

We will be collecting the data from multiple publicly available data sources.
This includes data about the CTA trips taken in the city of Chicago, all the bus routes, station entries about the riders count, daily boardings and totals, data about the weather, dates of public holidays in the country, crimes that have occurred etc.

DATA PREPARATION

1.Tasks of cleaning, selecting and formatting will ensure the Dataset is in usable format.
2.Eliminating redundant data and managing incomplete data.
3.Ensuring the same format for all the data columns to maintain consistency in the dataset.

DATA ANALYSIS

To analyze the CTA data :

1. Preparing a report outlining the predictive methodology used to build the model.
2. Implementing regression techniques to form the CTA optimal route model.
3. Feature selection.
4. Evaluate the model's performance by using measures such as accuracy, recall and F-score.
5. Plotting the data (rides vs. weather, rides vs. time, rides vs. crimes, rides vs. weather, etc) to give visual impact.

PROJECT OUTLINE:

LITERATURE REVIEW:

- "Spatial Analysis of Crime Incidents in Chicago" by Amanda L. Fath (2018)": A study using GIS techniques to analyze crime incident spatial patterns in Chicago- insights into impact of environmental factors on the crime rates.

- "Examining the Relationship Between Public Transit Use and Crime" by Jeffrey R. Lin (2017): Research exploring the association between public transit usage and crime rates.

DATASET SOURCE:

- Historical ridership data from Chicago Transit Authority ranging from 2013 to 2023 These datasets were accessed through the City of Chicago Data portal.
  https://data.cityofchicago.org/Transportation/CTA-Ridership-Daily-Boarding-Totals/6iiy-9s97/about_data

  https://data.cityofchicago.org/Transportation/CTA-Ridership-L-Station-Entries-Monthly-Day-Type-A/t2rn-p8d7/about_data

  https://data.cityofchicago.org/Transportation/CTA-Ridership-L-Station-Entries-Daily-Totals/5neh-572f/about_data

  https://data.cityofchicago.org/Transportation/CTA-Ridership-Bus-Routes-Monthly-Day-Type-Averages/bynn-gwxy/about_data

- Crime data spanning from 2001 to the present in Chicago was sourced from the City of Chicago Data portal.
  https://data.cityofchicago.org/Public-Safety/CTA-Crime/5xiy-qnsz

- Historical weather data for Chicago was retrieved from the National Centers for Environmental Information.
  https://www.ncei.noaa.gov/access/past-weather/chicago

- US Holidays dates from 2004 to 2023 were gathered from a dataset titled 'US Holiday Dates'.
  https://holidayapi.com/countries/us-il/2023

- Seasonal data is also included in the dataset and has been hard-coded to ensure stability as it does not undergo regular changes.

DATASET DESCRIPTION:

CTA RIDERSHIP DATA (Bus Routes)

FIELD	**DESCRIPTION**
Avg_Saturday_Rides	Average number of rides on weekends (Numeric)
Avg_Sunday_Holiday_Rides	Average number of rides on Sunday Holidays (Numeric)
Avg_Weekday_Rides	Average number of rides on weekdays (Numeric)
Monthtotal	Cumulative count of rides for the month (Numeric)
Month_Beginning	Beginning of month (Date & Time)
Route	Assigned route number (Numeric)
Route Name	Street name served by the route (Text)

CTA RIDERSHIP DATA (Daily ‘L’ Station Entries)

FIELD	**DESCRIPTION**
Date	Date corresponding to recorded data *Date & Time)
Daytype	Type of day (Text)
Rides	Number of rides recorded on that day (Numeric)
Station_id	Unique ID assigned to boarding station (Numeric)
Stationname	Station name or street name where station is located (Text)

CTA RIDERSHIP DATA (Daily Boarding Total)

FIELD	DESCRIPTION
Bus	Total Count of Buses servicing on the day (Numeric)
Day_type	Type of the day (Text)
Rail_Boardings	Total count of people boarding trains (Numeric)
Service_date	Date of service (Date & Time)
Total_rides	Number of rides in a day (Numeric)

CTA CRIME DATA

FIELD	DESCRIPTION
Arrest	Specifies if an arrest was made (Checkbox)
Beat	Geographical area where incidence occurred (Text)
Block	Address of incidence (Text)
Case Number	CPD Records Division No. unique to incidence (Text)
Community	Specifies the community area of incidence (Text)
Date	Date of incidence, can also be best estimate (Date & Time)
Description	Secondary description of IUCR code. Also a subset of primary description. (Text)
Domestic	Specifies if if incidence is domestic-related by IDVA (Checkbox)
ID	Unique ID for every record
IUCR	IUCR(Text) codehttps://data.cityofchicago.org/d/c7ck-4 38e.
Location Description	Description of location of incidence (Text)
Primary Type	Primary description of IUCR (Text)
Ward	Ward where incidence occurred (Text)

DATA FOR HOLIDAYS 2023

FIELD	**DESCRIPTION**
Date	Date of the holiday (Date & Time)
Name	Holiday name (Text)
Notes	Additional Info (Text)
WeekDay	Day of week (Text)

DATA - CHICAGO TEMPERATURE

FIELD	**DESCRIPTION**
Day	Date corresponding to weather (Date & Time)
Month	Month corresponding to data (Date & Time)
Year	Year corresponding to data (Date & Time)
Precipitation	Amount of rain, ice pellets and snow on ground. (Numeric)
Temperature	Temp of day based on Minimum, Maximum or observed (Numeric)

DATA PROCESSING:

- Integrate all data set csv files in one to easily access the data.
- Identify and eliminate all empty or null values from the dataset to ensure data completeness.
- Unify date and time formats for enhanced uniformity.
- Standardize column data formats.
- Implement various cleaning techniques to detect and remove duplicates or anomalies.
- Employ outlier detection mechanisms to address unusual data points.

MODEL SELECTION:

Feature Selection - LASSO Regression:

- Given the data set's extensive feature sets, the initial step involves applying LASSO regression to assess the significant association between characteristics and output variables.
- Subsequently, the final model can be trained using this selected set of predictors.

Generalized Additive Models for Inference & Predictors, Piecewise Polynomial Regression & Random Forest:

- GAM is straightforward and it offers flexible prediction functions to uncover underlying data patterns. Regularization of predictor functions also assists in preventing overfitting.
- To explore ride counts and their changes during COVID period, we'll use piecewise polynomial regression, aiming to visualize the trend and observe the return to normal behavior.
- GAM may produce odd outcomes based on input predictors, hence we plan to compare it with random forest. Despite being unable to adjust the smoothness of predictor functions in GAM, random forest can highlight discrepancies if the GAM model proves to be significantly inaccurate because it is more of a "Black Box" technique.

METRICS

The assessment of output results involves application of different metrics such as:

- MSE, RMSE and R2 - These are used to quantify the variance between predictors and responses.
- F-Statistics, P-value, VIF and RSE - These are used to determine optimal models.
- The accuracy measure will be applied to evaluate the correctness of the model.

TOOLS

- We will be using programming languages such as Python and R for data analysis and statistics for our project. Libraries such as dplyr, randomForest, gam, matplotlib, scikit-learn, Pandas, Numpy, ggplot2, seaborn etc will be imported and utilized.
- Jupyter Notebook to create code, equations and visualizations.
- R Studio which is a user-friendly interface for programming R.
- Tableau which is a data visualization tool used to create interactive and shareable dashboards.
- GitHub is used for managing and sharing code repositories.

REFERENCES

1. "CTA-Annual Ridership Report" Chicago Transit Authority Assets report (Jan24,2022).
2. "Transit Trends - Chicago Metropolitan Agency for Planning" CMAP analysis of Federal Transit Administration. National Transit Database. Percent change in vehicle speed, 2005-15.
3. " Mass Transit Security in Chicago" RC Johnson.
4. https://data.cityofchicago.org/Public-Safety/Crimes-2022/9hwr-2zxp/data
5. "https://www.researchgate.net/publication/348251988_Behavioral_dynamics_of_public_transit_ridership_in_Chicago_and_impacts_of_COVID-19 "Fissinger, Mary. (2020). Behavioral dynamics of public transit ridership in Chicago and impacts of COVID-19.
6. https://www.researchgate.net/publication/335270812_Preferences_for_travel-based_multitasking_Evidence_from_a_survey_among_public_transit_users_in_the_Chicago_metropolitan_area Krueger, Rico & Rashidi, Taha & Auld, Joshua. (2019). Preferences for travel-based multitasking: 65. 334-343. 10.1016/j.trf.2019.08.004.