CSP 571 - Data Preparation and Analysis
Project Proposal

Title:

Chicago Transit Authority Data Analysis

Team Members

Pranit Kotkar
pkotkar1@hawk.iit.edu
A20512027

Siddhi Shukla
sshukla12@hawk.iit.edu
A20516414

Anushka Chaubal
achaubal@hawk.iit.edu
A20511568

Rewa Deshpande
rdeshpande1@hawk.iit.edu
A20492328

Vaishnavi Shankar Devadig
vdevadig@hawk.iit.edu
A20516246

Chicago Transit Authority Data Analysis

Connecting Chicago, One Ride at a Time

ABSTRACT:

The goal of this research is to use data analysis to determine the optimal transit routes in Chicago based on the weather, crime, and COVID-19 epidemic. The research will examine data from numerous sources, including passenger counts, ticket sales, railway and bus timetables, weather, crime, and COVID-19 databases, using data analytic techniques like regression analysis, time series analysis, and data visualization. The research will determine which transit routes are the most effective and secure for CTA users and offer suggestions on how CTA may enhance its services based on the data analysis. The project's intended results include identifying changes in transportation system usage patterns brought on by weather, crime, and the COVID-19 pandemic as well as offering suggestions for how CTA should modify its offerings to better serve its customers. The project will increase access to dependable, safe, and efficient public transit, which will benefit CTA passengers as well as the larger Chicago community.

RESEARCH GOAL:

Our objective of this study is to determine the optimum transit routes for Chicago Transit Authority (CTA) passengers by analyzing data from a variety of sources, including COVID-19 statistics, weather, crime, and other datasets. In order to determine the most effective and secure transit routes for CTA consumers, our research compares data from several datasets using data analysis techniques like regression analysis, time series analysis, and data visualization. Additionally, our research intends to make suggestions for how CTA might enhance its services in light of data analysis and modify those services to better suit evolving consumer demands. Customers of CTA and the greater Chicago community will gain from the research by having easier access to dependable, safe, and efficient public transportation services.

WHAT WE AIM TO ADDRESS:

- What are the transit routes with the highest number of boarding passengers?
- Compare the number of trips on various routes between weekdays and weekends.
- The impact of seasons and temperature on the usage of Chicago Transit Authority.
- The impact of holidays and weekends on the usage of Chicago Transit Authority.
- The influence of crime statistics in the particular area and route number on the number of trips held.
- How drastically have the passenger statistics changed due to COVID-19?
- To provide recommendations to CTA based on which of the routes have a higher number of passengers but not enough transit vehicles to satisfy the needs, the crime statistics and how the problem can possibly be alleviated.
- Location density diagrams for the busiest stations.
- How can the number of travels in a certain period of time be predicted?
- Analyzing gender-specific data for each of these routes.

PROPOSED METHODOLOGY:

Brief description of proposed approach that we will be implementing to answer above mentioned questions:

ASSUMPTIONS:

- Considering data for the time period of 2013-2022.
- Weather will be uniform throughout the United States.

DATA COLLECTION:

● Historical CTA data (2013-2022) collected from CTA Trips | City of Chicago | Data Portal

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

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

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

https://data.cityofchicago.org/Transportation/CTA-Ridership-Daily-Boarding-Totals/6iiy-9s97

● Weather data collected from National Centers for Environmental Information

https://www.ncei.noaa.gov/access/past-weather/chicago

● Holidays - US Holiday Dates (2004-2021) collected from US Holiday Dates (2004 - 2021)

https://holidayapi.com/countries/us-il/2023

● Seasonal data - Can be hard coded as they don't change for a country

● Crime data collected from Crimes - 2001 to Present | City of Chicago | Data Portal

https://data.cityofchicago.org/Public-Safety/CTA-Crime/5xiy-qnsz

DATA PREPARATION:

● Simplifying the dataset by performing tasks such as cleaning, selecting, and formatting features.

● Remove any data that is redundant, incomplete, or unnecessary.

● Additionally, the format of the data columns may need to be adjusted to ensure consistency throughout the dataset. For instance, if the dates in some records are in DD-MM-YYYY format and in others they are in MM-DD-YY format, all dates will be modified to the MM-DD-YYYY format.

DATA ANALYSIS:

The descriptive and prescriptive techniques listed below will be used to analyze the CTA data :-

1. Provide a clear report on the predictive methodology that was used to build the model.

2. Implementing regression techniques to construct the CTA best route model.

3. Feature selection.

4. Using the accuracy, recall, and f score measures to assess the model's performance.

5. Plot the data (e.g., rides vs. weather, rides vs. COVID-19, rides vs. time, rides vs. crimes, rides vs. weather, etc.).

PROJECT OUTLINE:

DATA PROCESSING:

- Find and remove all empty values from each dataset.
- Make necessary changes to the column data format to provide uniformity across all datasets. (Example: Date and time format)
- One file created by combining multiple csv files of a data source.
- If there are any duplicates or anomalies in the datasets, find them and remove them.

DATASET DESCRIPTION:

CTA - Ridership - Bus Routes - Monthly Day-Type Averages & Totals

Sl No.	Column	Data type	Description
1	Route	Number	The route number assigned.
2	Route Name	Plain Text	The name of the street the route serves.
3	Month_Beginning	Date & Time	The beginning of a month
4	Avg_Weekday_Rides	Number	The average number of rides on weekdays.
5	Avg_Saturday_Rides	Number	The average number of rides on weekends.
6	Avg_Sunday_Holiday_Rides	Number	The average number of rides on Sunday Holidays
7	Monthtotal	Number	The total number of rides in a month.

CTA - Ridership - 'L' Station Entries - Daily Totals

Sl No.	Column	Data type	Description
1	Station_id	Number	The ID assigned to a boarding station.
2	Stationname	Plain Text	The name of the station or the street the station is in.
3	Date	Date & Time	The date corresponding to the data.
4	Daytype	Plain text	The type of day.
5	Rides	Number	The number of rides on that day.

CTA - Ridership - 'L' Station Entries - Monthly Day-Type Averages & Totals

Sl No.	Column	Data type	Description
1	Station_id	Number	The ID assigned to a boarding station.
2	Stationname	Text	The name of the station or the street the station is in.
3	Month_Beginning	Date and time	The beginning of a month

4	Avg_Weekday_Rides	Number	The average number of rides on weekdays.
5	Avg_Saturday_Rides	Number	The average number of rides on weekends.
6	Avg_Sunday_Holiday_Rides	Number	The average number of rides on Sunday Holidays
7	Monthtotal	Number	The total number of rides in a month.

CTA - Ridership - Daily Boarding Totals

Sl No.	**Column**	**Data type**	**Description**
1	Service_date	Date & Time	The date of the service.
2	Day_type	Plain Text	The type of day.
3	Bus	Number	The number of buses servicing on that day.
4	Rail_boardings	Number	The number of people boarding trains.
5	Total_rides	Number	The number of rides on that day.

CTA - Crime Data

SR. No	Column Name	Type	Description
1	ID	Number	Unique Identifier for each record
2	Case Number	Plain Text	The Chicago Police Department RD Number (Records Division Number), which is unique to the incident.
3	Date	Date and Time	Date when the incident occurred. This is sometimes a best estimate.
4	Block	Plain Text	The partially redacted address where the incident occurred, placing it on the same block as the actual address.
5	IUCR	Plain Text	The Illinois Unifrom Crime Reporting code. This is directly linked to the Primary Type and Description. See the list of IUCR codes at https://data.cityofchicago.org/d/c7ck-4 38e.
6	Primary Type	Plain Text	The primary description of the IUCR code.
7	Description	Plain Text	The secondary description of the IUCR code, a subcategory of the primary description.
8	Location Description	Plain Text	Description of the location where the incident occurred.

9	Arrest	Checkbox	Indicates whether an arrest was made.
10	Domestic	Checkbox	Indicates whether the incident was domestic-related as defined by the Illinois Domestic Violence Act.
11	Beat	Plain Text	Indicates the beat where the incident occurred. A beat is the smallest police geographic area – each beat has a dedicated police beat car. Three to five beats make up a police sector,
12	District	Plain Text	Indicates the police district where the incident occurred.
13	Ward	Plain Text	The ward (City Council district) where the incident occurred.
14	Community	Plain Text	Indicates the community area where the incident occurred.

Chicago Temperature Data:

SR. No.	Column	Data type	Description
1	Year	Date & Time	The year corresponding to the weather data.
2	Month	Date & Time	The month corresponding to the weather data.
3	Day	Date & TIme	The date corresponding to the weather data.
4	Temperature (F)	Number	Has three sub columns: · Minimum temperature of the day. · Maximum temperature. · Observed temperature.
5	Precipitation	Number	The amount of rain, melted snow, snow, or ice pellets on the ground.

Holiday data for 2023:

SR. No.	Column	Data Type	Description
1	Date	Date & Time	Date of the holiday.
2	Weekday	Plain text	Day of the week.
3	Name	Plain Text	Name of the holiday.
4	Notes	Plain Text	Additional significance.

MODEL SELECTION

Lasso Regression for feature selection

- As the above datasets contain a large number of features, we will first run a Lasso regression on them to determine whether characteristics have a significant link with our output variable.
- We may then train our final model using this collection of predictors.

GAM (Generalized additive models) for Inference and Predictions, Piecewise polynomial regression and Random forest.

- Interpreting GAM is simple. We are given adaptable prediction functions that we may use to find underlying patterns in the data. The predictor functions' regularization would also aid in preventing overfitting.
- We will perform a piecewise polynomial regression because we think it will help us see the ride counts, how they changed over the covid period, and the graph showing the return to normal behavior.
- While the GAM may provide odd outcomes depending on the input predictors, we intend to compare it with Random forest. Although we are unable to change the predictor functions' level of smoothness, random forest is more of a "black box" technique that would alert us if the GAM model was really wrong.

TOOLS:

- Programming Language - Python, R
- Software - Jupyter Notebook, R Studio, Tableau
- Libraries/Packages - dplyr, randomForest, gam, ggplot2, lm, scikit-learn, pandas, NumPy, seaborn, matplotlib
- Project Management and Source Control - GitHub

METRICS TO MEASURE ANALYSIS RESULTS:

The output results can be evaluated using these measures:

- The variation between predictors and answers is calculated using the MSE, RMSE, and R2 measures.
- RSE, VIF, F-statistics, and P-value are used to determine the optimal model.
- The accuracy measure will be used to assess the model's correctness.

REFERENCES:

- https://www.axios.com/local/chicago/2022/04/21/mapping-cta-crime-statistics-chicago-trains
- https://www.nctr.usf.edu/pdf/527-14.pdf
- https://www.wsdot.wa.gov/partners/erp/background/ST3%20Draft%20RidershipForecastingMethodologyReport_6March2015.pdf
- https://www.apta.com/research-technical-resources/research-reports/transit-workforce-shortage/
- https://www.researchgate.net/publication/354472520_Examination_of_New_York_City_Transit's_Bus_and_Subway_Ridership_Trends_During_the_COVID-19_Pandemic