GIS 5572 - Semester Project
Title: Traffic Impact Prediction Model

Name 1: Laure Briol
Name 2: Logan Gall
Name 3: Greg Kohler
Date: 4/11/2024
Project Github repository of all code: https://github.com/logan-gall/TrafficImpactPrediction

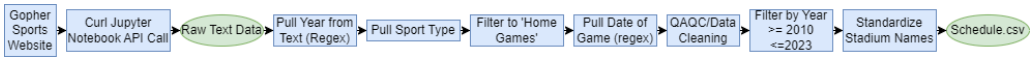
Abstract
This project aims to predict traffic impacts due to sporting events near the UMN campus. We created historical road network analysis layers for 500 sporting events utilizing historical traffic data. Three linear regression models were made to predict traffic impacts, but variability of large events are difficult to model fully.

Problem Statement
Large sporting events lead to substantial traffic impacts. We used a stacked approach, utilizing open source softwares, google cloud platform, and ESRI tools to create a traffic impact model. ArcGIS Pro Notebooks was used to build ETL pipelines that gather event schedules and road network data. We developed three machine learning models that predicted traffic cost and sent results to a Google Cloud PostGIS database. Access is provided through a Flask API deployed on Google Cloud Run, built from a GitHub repository. This API can be added in ArcGIS online web maps, allowing for a display of predicted traffic impacts.

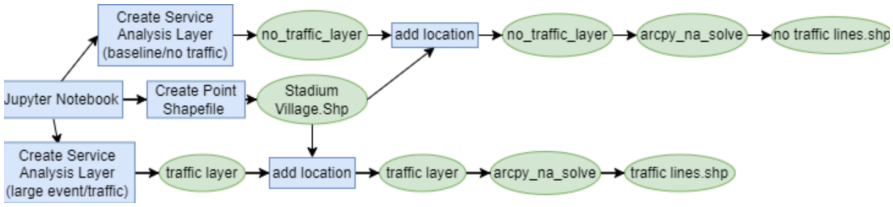
Input Data

#	Title	Purpose in Analysis	Link to Source
1	Road Network Analysis Lines	Road network analysis lines using historical data from ESRI's World Traffic Service	https://arcg.is/v8P9H
2	Gopher Sports Schedules	Large sporting events schedule from the University of Minnesota	Gopher Sports Main Page Text Schedule Example

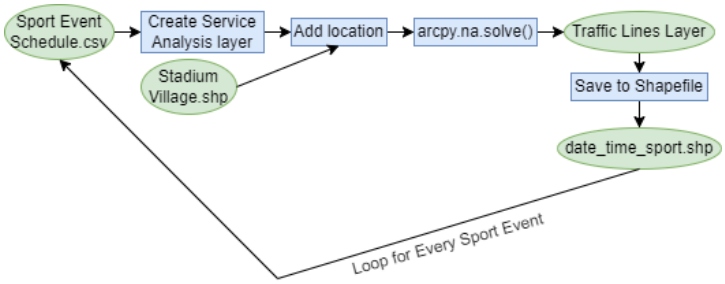
Data Flow Diagram(s) for System (in order)
Get Sports Schedule Diagram:



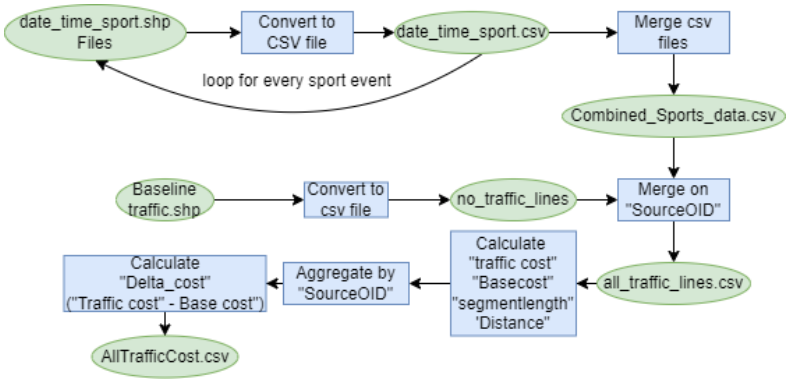
RoadlinesQA Diagram:



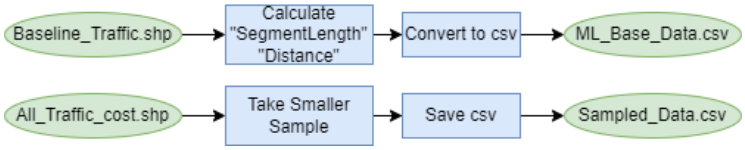
RoadLoop Diagram:



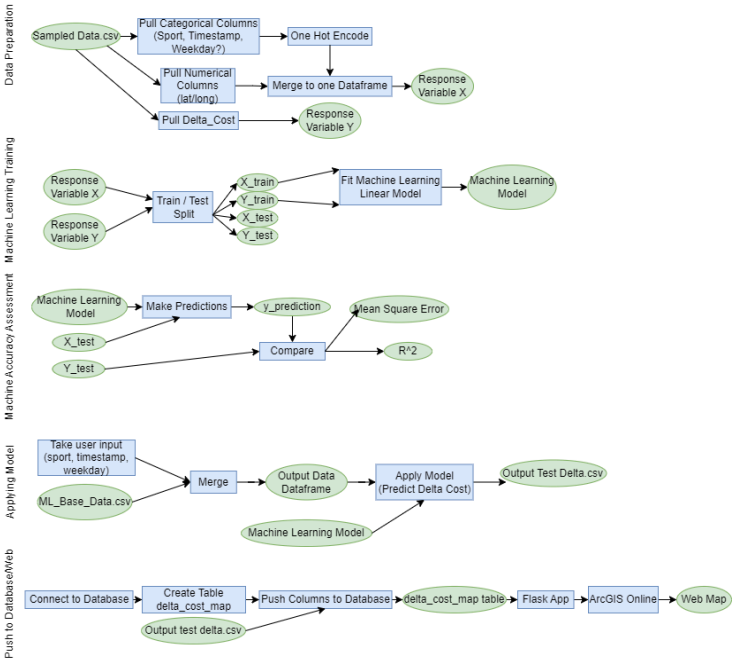
Data Cleaning1 Diagram:



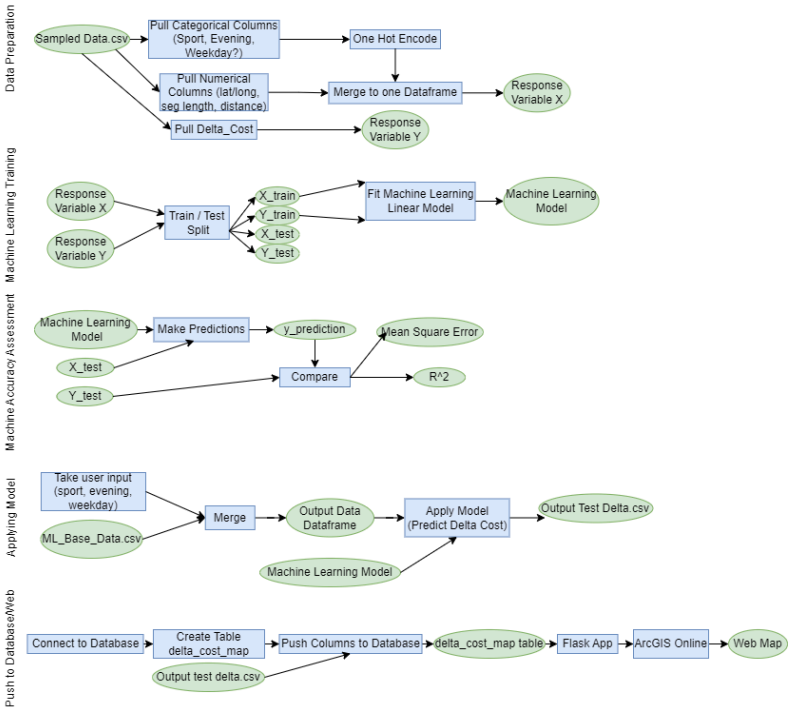
Data Cleaning 2 Diagram:



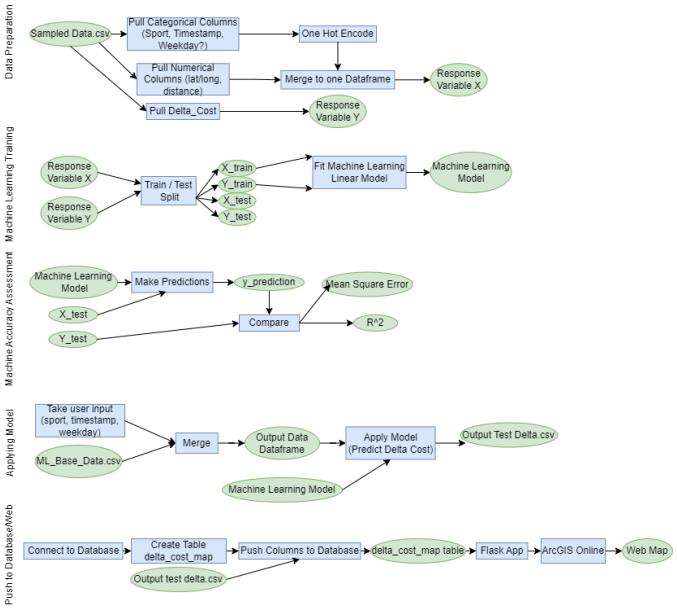
Machine Learning 1 Diagram:



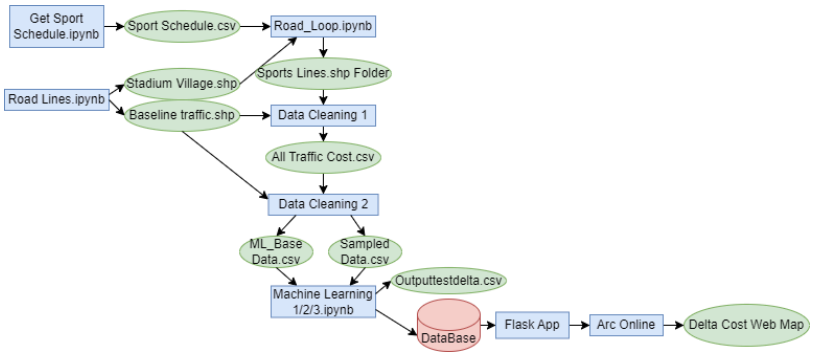
Machine Learning 2 Diagram:



Machine Learning 3 Diagram:



Overall Structure Diagram:



Model Comparison

#	Model Name	Evaluation Metric 1 (R^2 Score)	Evaluation Metric 2 (Mean Squared Error)	Rank Score of Model
1	5 Variable (Machine Learning 1)	$R^2 = 0.0019$	MSE = 0.03234	3
2	7 Variable (Machine Learning 2)	$R^2 = 0.0708$	MSE = 0.03011	1
3	6 Variable (Machine Learning 3)	$R^2 = 0.0107$	MSE = 0.03206	2

5 Variable

This model is a machine learning linear regression model using a mix of categorical and numerical variables.
The categorical input variables are:

- “Sport” (Type of sporting event)
- “Timestamp” (Time at which the event took place)
- “Weekday?” (Indicator True/False if the event occurred on a weekday Monday-Thursday)

The numerical input variables are:

- “Latitude” (Centroid Latitude of road segment)
- “Longitude” (Centroid Longitude of road segment)

7 Variable
This model is a machine learning linear regression model using a mix of categorical and numerical variables.
The categorical input variables are:

- “Sport” (Type of sporting event)
- “Evening?” (Indicator True/False if the event occurred during the evening after 3:01pm)
- “Weekday?” (Indicator True/False if the event occurred on a weekday Monday-Thursday)

The numerical input variables are:

- “Latitude” (Centroid Latitude of road segment)
- “Longitude” (Centroid Longitude of road segment)
- “Segment Length” (Length of road segment)
- “Distance” (Distance from the stadium village road network analysis starting point)

6 Variable
This model is a machine learning linear regression model using a mix of categorical and numerical variables.
The categorical input variables are:

- “Sport” (Type of sporting event)
- “Timestamp” (Time at which the event took place)
- “Weekday?” (Indicator True/False if the event occurred on a weekday Monday-Thursday)

The numerical input variables are:

- “Latitude” (Centroid Latitude of road segment)
- “Longitude” (Centroid Longitude of road segment)
- “Distance” (Distance from the stadium village road network analysis starting point)

Recommendation for Decision-making
The model recommended to be used for decision making for the traffic impact prediction would be the seven variable model. The seven variable model is the second machine learning notebook and has, of all the models, the highest overall R^2 score paired with the lowest Mean Squared Error (MSE) score.

Reflection: What did you learn? What would you do differently if you did this again?

This project has been a learning experience for everyone involved. One of the biggest things learned early on in the project is how vital data cleaning is to creating a good end product. When gathering sporting event data, several rows had to be cleaned in ways not previously considered. This included finding the correct year for every event, finding the right sport name, and formatting the time correctly. When starting the project, we did not factor in how much time data cleaning alone takes. If we had to do it again, we would give ourselves a longer period of time to make sure the data was properly cleaned and ready for analysis.

Another lesson learned was the importance of keeping well organized code. When working in a group, it is easy to have many different versions of code that do similar things. Through this process, we learned the importance of a shared repository and the advantages of being able to merge our changes. It is also important to name files in an easy to track way. If we had to do it again, we would decide on which notebooks we needed and their names before we started the process of writing code.

On the machine learning side, we also learned how different inputs can impact a prediction and lead to a model being more realistic for the given situation, even with several inputs. For example, it was informative to see how adding in distance from the stadium could dramatically change the way the model looks. Every factor has a role to play in prediction. If we could do it again, we would have considered more factors to go into the model earlier on, such as if the game was a playoff or weather conditions on the day. This would give us different ways to predict traffic for the given event.

References

GopherSports.com. University of Minnesota Athletics, gophersports.com/.

Esri. "Routing." ArcGIS Developers, Esri, n.d., https://developers.arcgis.com/documentation/mapping-apis-and-services/routing/.

Esri. "World Topographic Map." ArcGIS Online, Esri, n.d., https://www.arcgis.com/home/item.html?id=ff11eb5b930b4fabba15c47feb130de4.

OpenAI. ChatGPT. OpenAI, Version 3.5, 2022, https://openai.com/chatgpt/.

Appendix

Mock user statement:

My name is Thomas and I work for the Minnesota Department of Transportation. I'm reaching out for you to help us build a pipeline to evaluate traffic impacts due to large events. They cause a lot of inefficiency in our roadways. The online system needs to show a map with live conditions and areas of high potential traffic in the future. Can you help us with this?

Also, we have some other developers that want to get the results in coded format. They can receive geojson. Can you do that?

Traffic Impact Prediction Specification

Overview

Objectives

Develop a system that:

- Predicts traffic congestion caused by large events.
- Provides information to users about potential traffic events.
- Integrates with existing traffic visualizations.
- Allows users to customize their based on location, event type, and severity of traffic.
- Provides an API service for data requests.

Problems:

- Little predictive analysis: Current consumer systems lack the ability to predict traffic caused by large events.
- Reactive systems: Existing traffic controls are mostly reactive to live conditions.
- Limited: Users can't easily preview traffic conditions according to their specific needs and preferences.
- Significant impacts: The impact of large events on local traffic is often underestimated, leading to insufficient planning and management.

Who It Directly Affects:

- Commuters and residents: Those who travel regularly and live close to event locations.
- Attendees: People attending the events who need information for better travel planning.
- Traffic controllers: Government and private entities responsible for traffic management.
- Organizers: Those organizing large events can use traffic flow information for better coordination.

Why it is Important Solve for MnDOT:

- Efficient management: Proactively managing traffic flow, especially during large events, can significantly reduce congestion and improve commute times.
- User experience: Providing customized traffic updates will improve the overall experience for commuters and event attendees.
- Resource allocation: Traffic authorities can allocate necessary resources based on predicted impacts.

- Data-driven planning: Organizers can use traffic predictions for more efficient planning of events, leading to better safety and smooth operations.

Motivation

Traffic is a major inefficiency in transportation that wastes time and causes excess pollution. While daily traffic is expected and common, large events don't happen every day and the traffic impacts are difficult to plan around. A better way to predict traffic conditions can be useful for many to optimize hassle and time.

Our expected outcome is to build a practical user web service. The service will allow users to view traffic conditions and prepare for travel based on upcoming events. Local residents, event organizers, and passing commuters can plan their travel knowing the traffic impacts of an upcoming event.

Definitions

Historic transit data: Live historical traffic conditions are kept for many years by ESRI and HERE. They show how much traffic impact may be at a certain road segment for a given date/time, back to around 2013.

Delta Cost: The difference in travel cost between road segments on the baseline traffic dataset and the large event dataset. This shows how much extra time it takes to travel across a given road segment during an event compared to baseline/clear road conditions.

Line Strings: A connection of points, displayed as a continuous line. In this traffic impact project, our road network data sets are displayed as multiple line strings.

Road Segment: A small chunk of road (line string). The road network dataset is split from one continuous line into multiple pieces/segments, so it is easy to perform analysis at a small spatial resolution.

Baseline Traffic: This is a road network dataset that is set to be a model for minimal traffic conditions. Picked to be very late at night in the middle of summer season. This road network dataset is used for comparison to any large event traffic.

Large Event Traffic: This is a road network dataset that estimates traffic conditions during a large event. This is compared to baseline traffic datasets to find extra travel time at different road segments.

Mean Square Error (MSE): The mean of the squared difference between a predicted value and its true value. Shows the distance and average accuracy of a machine learning model. A lower value means better model predictions.

R^2 : R-Squared / Coefficient of determination. This shows how much variance of the dataset is explained/correctly predicted by a model. A higher value up to 1 is better.

Ranked Model score: How our models rank in accuracy when compared to each other.

Machine Learning: Utilizing optimization techniques to have a computer automatically pick the best parameters for a given set of input data to predict a given output. For this traffic impact project, we use linear regression to predict the extra cost to travel a road segment.

Live Traffic Flow: Traffic flow is a measure of how fast vehicles are traveling across a certain area. If there is very slow movement of vehicles, the flow of traffic is low.

Large Events: Any event that has the potential to have a significant impact on traffic. For this traffic impact analysis, we are looking at Football, Basketball, Hockey, and Volleyball events on the UMN campus.

Model: An abstracted version of a real-world 'thing'. It is a way of trying to simplify and describe. For this traffic impact project, we use machine learning models to predict traffic at a given road segment.

Scope

Functional Requirements

Traffic Data

- ESRI World Traffic Service
    - Pulls data from HERE
- Live traffic flow
- Updates every 5 minutes
- Historical traffic flow
- API

Event Data

- Gopher Sports
- Essential -- a small area with many major events (UMN)
- Nice to have -- Entire TC area
- Optional -- Even larger area

Traffic prediction model

- Takes input of event location, time, and date

- Takes input of historical traffic on that date
- Outputs an area with high probability of slow traffic flow (linestring/polyline or other)
- Essential -- Outputs some area prediction of traffic time
- Nice to have -- Outputs specific roads with poor traffic (lines)
- Optional -- Outputs an entire area's traffic prediction, both good and bad roads (lines)

Web Map Interface

- Road network map
- Essential -- Highlights areas with poor traffic based on user input
- Nice to have/Optional -- Live data and predictions of when traffic will start to get bad

GeoJson API

- User input date/time/event of interest on local machine
- This runs and pushes data to google cloud
- Google Cloud Run Flask API Endpoint
- Returns areas with potential congestion

Non-Functional Requirements

- Easy to use interface
    - Essential -- Usable by someone with desktop software and knows how to read time-series traffic maps
    - Nice to have -- Usable by power users of typical map applications i.e. Google/Apple maps
- Low processing time
    - Essential -- < 10min for a new event area/request
    - Nice to have -- < 1 min
    - Optional -- Near instant for any upcoming events
- Reliable
    - Will be able to consistently predict bad traffic for a given location/event
    - Essential -- A few limited events (i.e. UMN sports)
    - Nice to have -- TC very large recurring events (i.e. professional sports)
    - Optional -- Smaller events (Shows/Plays, other events)

Out of Scope Requirements

- Any scale larger than Twin Cities (or even just UMN campuses)
- Predictive modeling for events that do not occur regularly
- Prescriptive analysis / giving optimal routes that avoid traffic

- Filtering out noise from regular daily traffic

Persona Acceptance Criteria

Who are the stakeholders impacted by the project's success? What are they trying to achieve?

As a developer I …

- Require access to APIs so that I can download the data and perform analysis
- Require a machine learning model so that I can adapt the traffic predictions based on event types

As an operator I…

- Require reliable data streams so that I can maintain database results without error
- Require a robust data flow documentation so that I can continue to provide high quality results

As an end user I

- Require a web interface so that I can utilize the data effectively
- Require regular updates so that I can view live data when planning events

Open Questions

- How do we go about making decent models for traffic flow?
    - How to handle input?
        - Traffic data
            - Line strings
        - Events occur over a period of time
            - Spatio-temporal analysis
            - Look at hour or two before and after
    - How to handle output?
        - Line string(s)
        - Delta Cost
- Gathering event information
    - API?
        - Limited to API events
    - Manual lists?
        - Significantly reduces scope
        - But ensures high data quality

Dependencies

Closed Source Dependencies:

- ESRI Traffic data
    - Built upon HERE API
- ESRI Infrastructure
    - ArcPro
    - ArcOnline
- Google Cloud
- GitHub
- UMN Gopher Sports data

Open Source Dependencies:

- Flask
- PostGIS

References

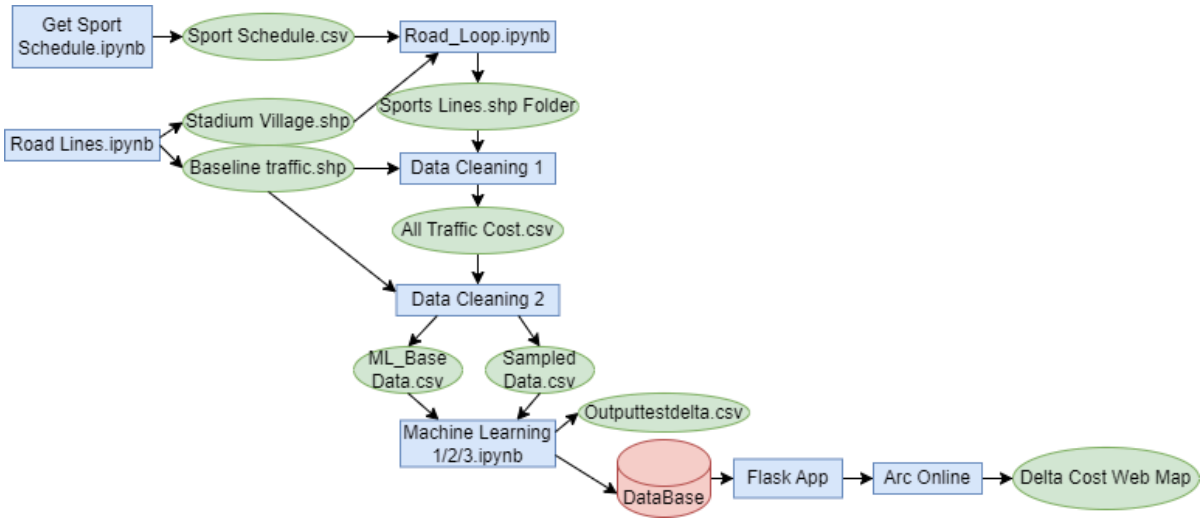
ESRI World Traffic Service:
https://umn.maps.arcgis.com/home/item.html?id=ff11eb5b930b4fabba15c47feb130de4

ESRI API Documentation:
https://developers.arcgis.com/documentation/mapping-apis-and-services/routing/

This subsection should provide a complete list of all documents referenced elsewhere in the SRS; This information may be provided by reference to an appendix or to another document.

Appendix

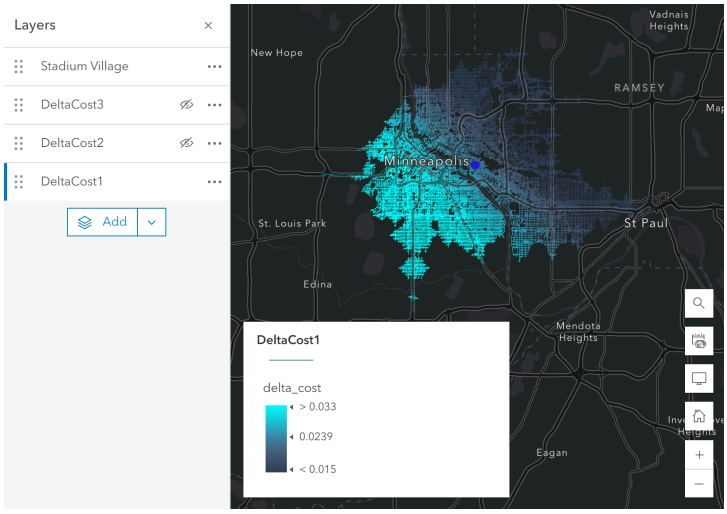
Traffic Impact Prediction Final Pipeline (requirement)



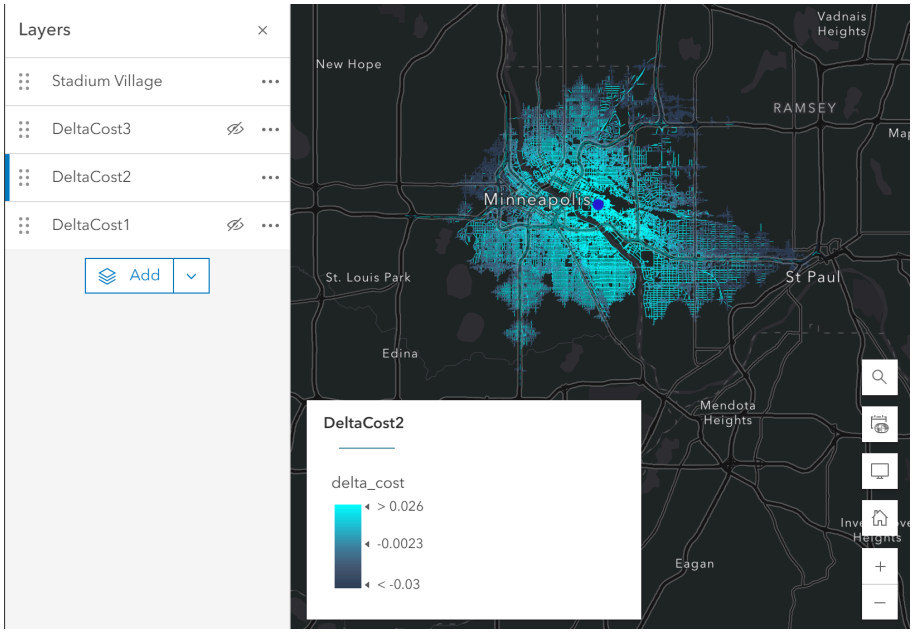
The structured pipeline provides stakeholders and project members with clear visibility into both the development process and final objective. Further, the Traffic Impact Prediction Final Pipeline promotes transparency and accountability by documenting each stage of the data processing and analysis workflow, enabling stakeholders to trace the lineage of predictions and insights generated by the system. The pipeline's modular design facilitates; scalability, flexibility, and maintainability, ensuring that the pipeline remains adaptable to emerging challenges and evolving needs within the realm of traffic prediction and management.

ArcGIS Online Final Map Outputs

Delta Cost Map for Machine Learning Model 1: Simulated Weekend Football Game at 7:00pm



Delta Cost Map for Machine Learning Model 2: Simulated Weekend Football Game at Evening:



Delta Cost Map for Machine Learning Model 3: Simulated Weekend Football Game at 7:00pm: