Rail Transit Delay Forecasting with Causal Machine Learning

Nishtha Srivastava*, Bhavesh N. Gohil*, and **Suprio Ray**#
*Sardar Vallabhbhai National Institute of Technology, India
#University of New Brunswick, Fredericton, Canada

STCausal @ SIGSPATIAL, 2024
Atlanta, GA, USA

Importance of rail travel

- Vital for the economy

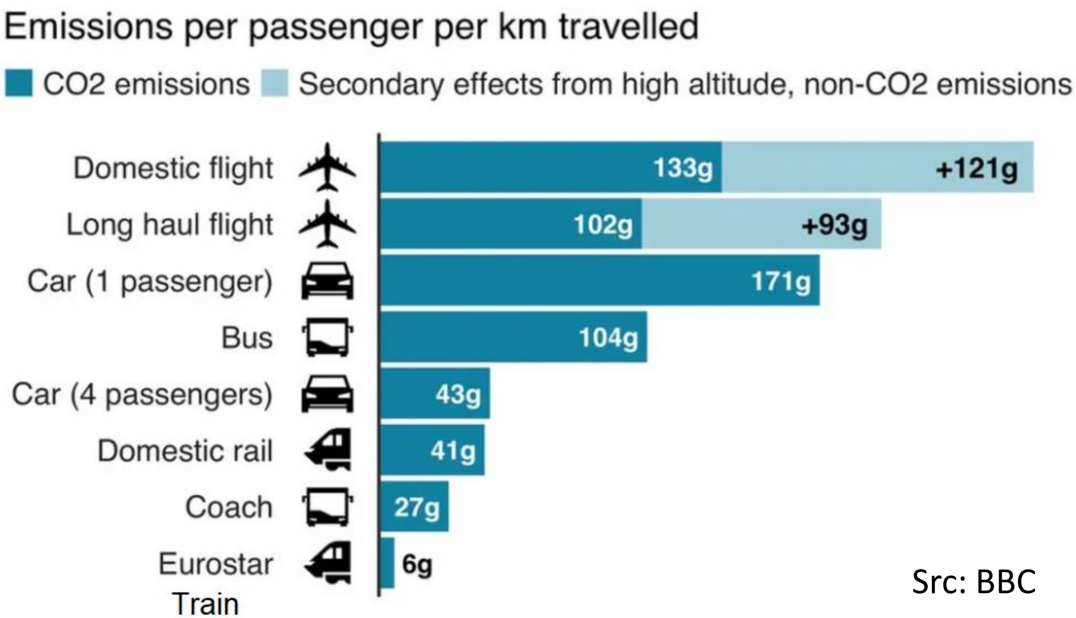
  - Indian rail network transports more than 11 billion passengers and 1.416 billion tons of freight annually

  - China's railways network delivered 3.66 billion passengers, and carried 4.389 billion tons of freight (2019)

- Comfortable and convenient
  - With private accommodations, onboard dining, and workspaces trains provide the ultimate comfort
  - Doesn't waste travel time waiting in line or going through security
  - Offers freedom to do our job or unwind while traveling by train

Importance of rail travel (cont.)

- Vital for the economy
- Comfortable and convenient
- Good for the environment



Src: BBC

Challenges faced by rail transit systems

- Rail transit systems face significant delays due to increased demand and capacity utilization

- This high utilization heightens the risk of delays propagating from one train to others, causing extended disruptions
  - Such delays can severely impact the efficiency and attractiveness of rail service

- Current delay prediction models are either event-driven or data-driven (Spanninger et al. 2022) but do not focus on causality

Our objective

- **Problem**: Current AI models, particularly DL models, operate as black-boxes
  - XAI methods like SHAP and LIME offer some insights into model predictions, but are limited and non-causal

- **Objective**: Improve rail transit delay predictions using causal machine learning techniques

Our contributions

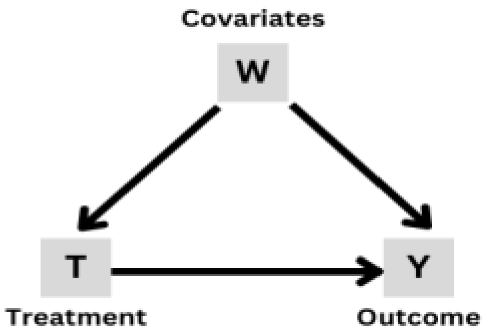
- Applying causal inference and ML techniques to forecast rail transit delays
- Using ITE and ATE to identify significant factors affecting rail transit delays
- Re-training ML models with identified key features and comparing their performance to the original models

Outline

- Motivation
- Background
- Our approach
- Evaluation and analysis
- Conclusion

Causal Inference

- Techniques to determine the impact of a treatment T on an outcome Y, considering covariates W, without requiring strict assumptions about the model's structure.



- **Treatment effect (T):** The change in the outcome variable resulting from a modification in the treatment variable.
- **Covariates (W):** Factors that influence both the treatment and the outcome.
- **Outcome (Y):** The resulting variable or output

Individual Treatment Effect (ITE)

- Let Y_1 be the outcome variable when the individual receives the treatment and Y_0 be the outcome variable when the individual does not receive the treatment.

- For an instance i with covariates X_i, its corresponding ITE is

$$ITE(X_i) = \mathbb{E}[Y_1|X_i] - \mathbb{E}[Y_0|X_i]$$

- E[Y_1 | X_i] represents the expected outcome Y_1 , given the individual's covariates X_i

- E[Y_0 | X_i] represents the expected outcome Y_0 , given the same covariates X_i

Average Treatment Effect (ATE)

- ATE for a given feature is determined by averaging the ITE values associated with that feature

- The ATE for a feature X_i can be computed as follows:

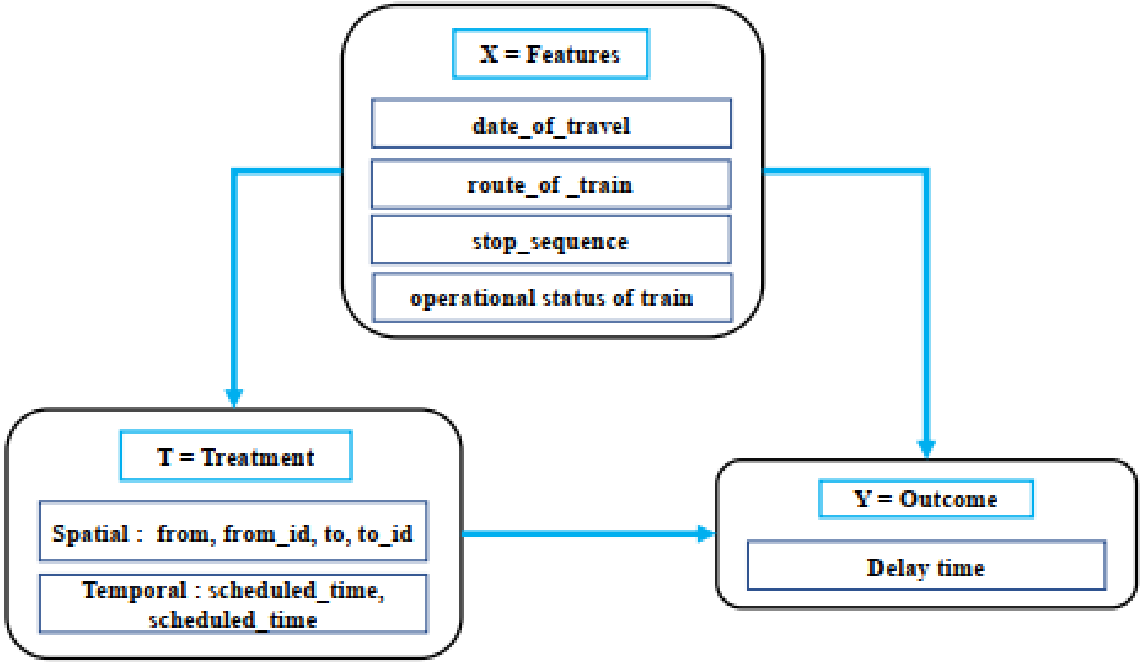
$$ATE = \mathbb{E}[Y_1 - Y_0]$$

Outline

- Motivation
- Background
- Our approach
- Evaluation and analysis
- Conclusion

Causal inference framework

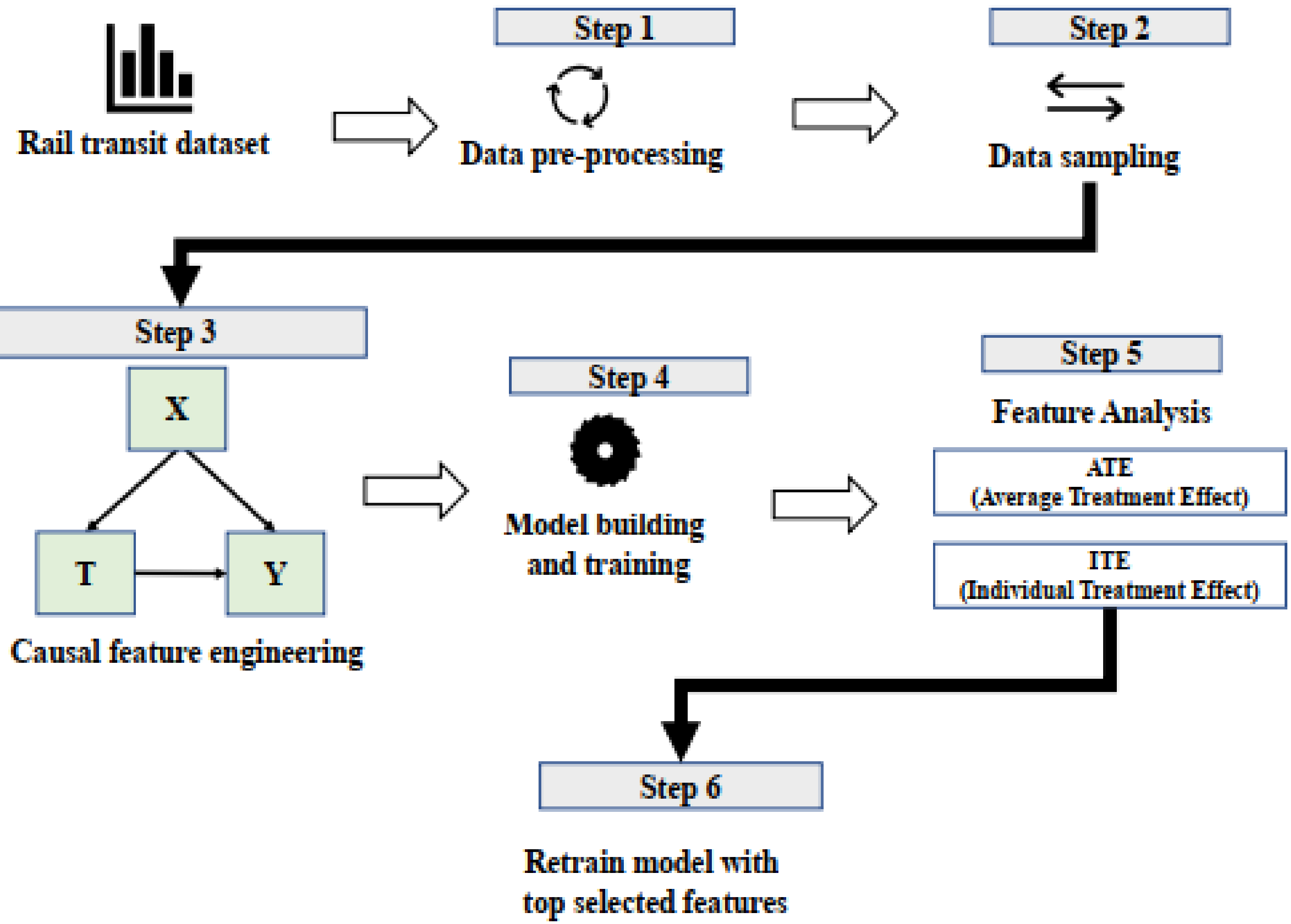
- The objective is to predict delays by assessing the causal effect of **interventions** (e.g., schedule changes, track maintenance) on the outcome (the delay), considering various **covariates** X (e.g., date_of_travel, route_of_train, stop sequence etc.)



Causal inference framework (cont.)

- **The outcome variable** $Y_i \in$ R represents the delay for rail transit unit i
- The **treatment variable** $T_i \in \{0, 1\}$ indicates whether a particular intervention or treatment was applied, where $T_i = 1$ if the intervention was applied and $T_i = 0$ otherwise
- The **covariates** $X_i \in$ R$_p$ form a vector of p observed features for rail transit unit i
  - When interventions are applied to features X_i in a causal model, if $T_i =$ 1 (indicating an intervention), the feature is adjusted to reflect the treatment's impact
  - For example, if X_i represents the train schedule and $T_i = 1$ corresponds to a **schedule change**, X_i would be replaced with the new schedule time. This can help estimate the treatment effect by comparing outcomes under treatment ($T_i = 1$) and no treatment ($T_i = 0$)

Overall approach



Steps in our approach

- **Step 1: Data Preprocessing**: Clean and preprocess the dataset to ensure consistency and quality: $D \leftarrow$ load_data("*rail_transit_dataset.csv*")
  Handle missing values and encode categorical features: $D' \leftarrow$ preprocess(D)

- **Step 2: Data Sampling**: Perform data sampling to ensure a balanced representation of treated and untreated units: $D_{\text{sampled}} \leftarrow$ sample_data(D')

- **Step 3: Causal Feature Engineering**: Identify and encode relevant features X that are hypothesized to influence the treatment T and outcome Y:
  $X_{\text{causal}} \leftarrow$ causal_feature_engineering(D_{sampled})

- **Step 4: Model Building and Training**: Train causal machine learning model to estimate ATE and ITE.
  Train the model using X_{causal}, Y, T: $\hat{f}(X, T) \leftarrow$ CausalModel(X_{causal}, Y, T)

- **Step 5: Feature Analysis**: Analyze and select the most significant features based on their estimated effects on Y:
  $F_{\text{top}} \leftarrow \{f \mid \text{ATE}(f) > \tau\}$

- **Step 6: Model Retraining**: Retrain the model using the top selected features to improve predictive performance: $\hat{f}_{\text{retrained}} \leftarrow$ CausalModel($X_{F\text{top}}, Y, T$)
  Evaluate the retrained model's performance and predict rail transit delay

Data Preprocessing

- Outlier detection and removal
  - Outliers are identified and removed to prevent their disproportionate influence on the model.

- Handling missing values
  - All missing values are categorized as "Unknown"

- Categorical variable transformation
  - The categorical attributes in the data are converted into a binary format using one-hot encoding

- Data integration.
  - The processed data is integrated into a unified dataset, ready for causal machine learning model training

Data sampling

- Time frame
  - Data from the dataset spanning the most recent two years were selected

- Data completeness
  - Only records with complete and valid entries were included. Missing or corrupted data entries were excluded to maintain the integrity of the analysis

- Operational metrics
  - Records including critical performance indicators such as delays, on-time performance, and train schedules were prioritized

- Geographic scope
  - Data from all relevant transit lines within the NJ Transit and Amtrak NEC networks were considered to provide a comprehensive view

Causal feature engineering

- Causal the features are grouped into three key components: X, Y, and T
  - X includes various covariates like weather conditions, address specifics, and vehicle attributes, offering essential contextual and environmental information
  - The Y component represents the outcome of interest, specifically train delays, which the causal analysis aims to understand or predict
  - The T element encompasses treatment variables, categorized into spatial (e.g. from, to) and temporal factors (e.g. scheduled time)

Model building and training

- After the data pre-processing stage, the dataset is trained using uplift tree classifier

- An uplift tree classifier is a decision tree designed to predict the incremental impact of an intervention or treatment
  - Finds the causal effects of a treatment by comparing potential responses of individuals if they get the treatment versus, if they do not

- It tries to determine the best split at each node by maximizing the difference in outcomes between the treatment and control groups

Feature analysis

- ATE estimates for various features impacting rail transit delays

Feature name	ATE values
stop_sequence	0.98
actual_time	0.97
scheduled_time	0.88
line	0.85
status	0.84
to	0.77
type	0.71
Date	0.70
to_id	0.66
from	0.64
from_id	0.61
train_id	0.59

Feature analysis (cont.)

- ITE of top 5 features

Sr. no.	Feature name	ITE key	ITE value
1	stop_sequence	1	0.77
		6	0.87
		18	0.98
		12	0.88
2	actual_time	02-04-2018 06:41	0.97
		02-03-2018 01:21	0.88
3	scheduled_time	02-04-2018 06:41	0.87
		07-04-2018 01:21	0.88
4	line	Northeast Corrdr	0.85
		Amtark	0.86
5	status	cancelled	0.88
		departed	0.84
		estimated	0.83

Retraining model

- **Feature changes**.
  - Rail transit systems often experience frequent changes, such as schedule updates, new routes, and alterations in operational procedures
  - It is essential to retrain the model with the most recent data to ensure it captures these changes effectively.

- **Concept drift**
  - The nature of delays and operational disruptions may evolve over time, leading to shifts in the underlying data distribution

Outline

- Motivation
- Background
- Our approach
- Evaluation and analysis
- Conclusion

Dataset overview

- NJ Transit Amtrak NEC Dataset* (1.5M records from 2018-2020)

- A trip transit dataset focusing on rail delays within the New Jersey Transit (NJ Transit) and Amtrak services along the Northeast Corridor (NEC)

*https://www.kaggle.com/datasets/pranavbadami/nj-transitamtrak-nec-performance

Dataset details

- NJ Transit Amtrak NEC Dataset

Attribute	Description	Example Record
Date	The date on which the train operation occurred.	2024-09-01
train id	A unique identifier for each train service.	63
stop sequence	The sequence number indicating the order of stops for a train.	3
from	The station where the train originates or departs from	Philadelphia
from id	unique identifier for the departure station.	1
to	The destination station where the train is scheduled to arrive.	Newark Airport
to id	A unique identifier for the arrival station.	37953
scheduled time	The planned departure or arrival time according to the schedule.	12:00
actual time	The actual recorded departure or arrival time	12:10
delay minutes	The time difference in minutes between the scheduled and actual time,representing the delay	10
status	The operational status of the train	Cancelled
line	The train line or route that the train is following	Northeast Corrdr
type	The type of train service	NJ Transit

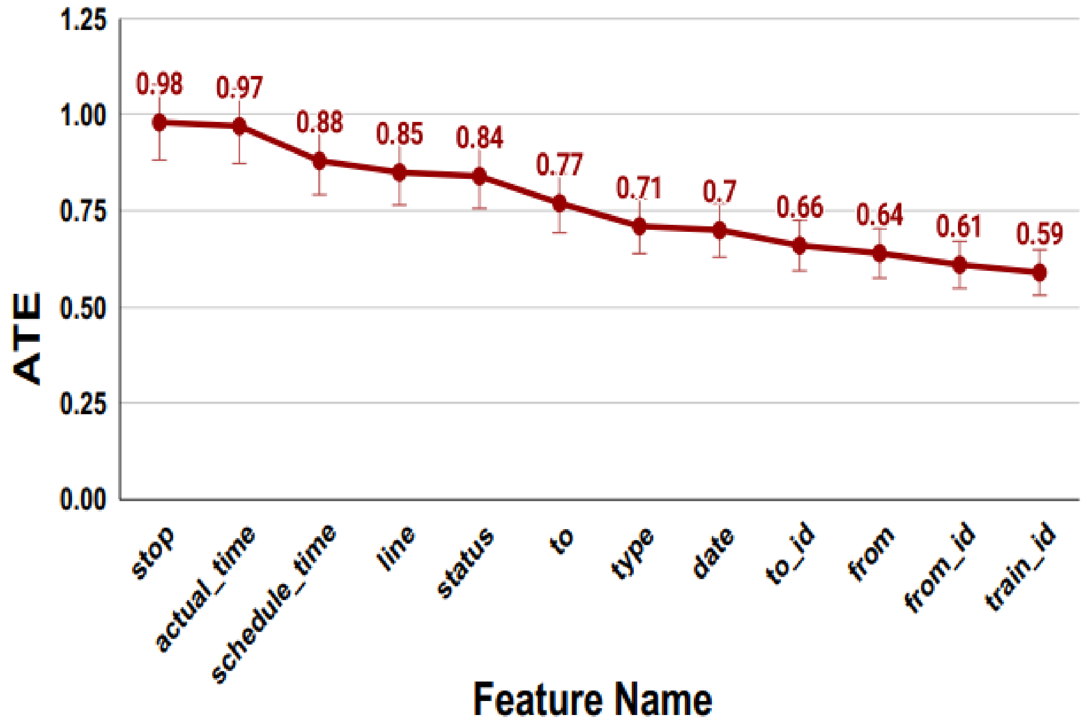
Experimental setup

- The experiments were conducted using Python 3.12.3 on a server equipped with a 3.31 GHz Intel(R) Xeon(R) CPU and 16 GB of RAM.

- We employed the CausalML library for implementing causal machine learning techniques.
  - This library offers various uplift modeling and causal inference methods, utilizing advanced algorithms based on contemporary research.

ATE values

- ATE values of the features in NJ transit dataset



Top ranked features

- Top ranked features of NJ transit dataset according to ATE score

Feature rank	Feature
1	stop_sequence
2	actual_time
3	scheduled_time
4	line
5	status

1. **stop_sequence (ATE: 0.98):** The order in which a train stops significantly affects the outcome. This is likely due to the cumulative effect of stopping at multiple stations impacting delays or performance.

2. **actual_time (ATE: 0.97):** The actual time at which a train departs or arrives is crucial in determining delays, as deviations from the schedule are directly observed here.

3. **scheduled_time (ATE: 0.88):** reflects the planned schedule's influence

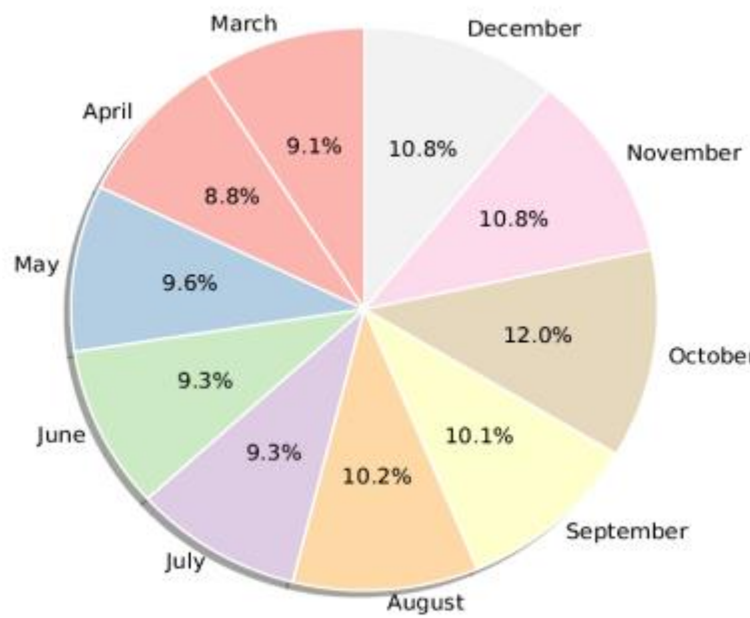
4. **line (ATE: 0.85):** Different train lines or routes have varying performance characteristics or constraints,

5. **status (ATE: 0.84):** The status of the train, whether it is "cancelled," "departed," or "estimated,"

Seasonal Trends

- October shows the highest percentage of delays (12%), likely due to weather conditions

- Winter months (December, November) also show elevated delays



UNB FREDERICTON

Comparison with baselines

- Accuracy comparison of classifiers with our Causal inference approach against different ML approaches

Accuracy without selecting features when trained using XGB	Accuracy without selected features when trained using RF	Accuracy without selecting features when trained using SVM	Accuracy with selected features when trained using UpliftTreeClassifier
93.4%	91.89%	90.56%	**95.6%**

Outline

- Motivation
- Background
- Our approach
- Evaluation and analysis
- Conclusion

Conclusion

- We have proposed a causal inference based framework to forecast rail transit delays
  - Utilize ML technique (uplift tree classifier) to predict the incremental impact of an intervention or treatment

- Our framework provides insights about rail transit delay
  - Stop sequence and actual time have the most significant impact on delays.
  - Scheduled time, line, and status also contribute to delays.
  - Delays are notably higher during peak times, suggesting a need for capacity adjustments.

Future work

- Focus on infrastructure and specific train characteristics to further improve transit delay predictions

UNB
FREDERICTON

Thanks!