Data Mining Transit Information from King St Pilot Project

Chantal Sylvestre

CKME136: Summer 2019

Project files: https://github.com/chantalsylvestre/CKME136Project

Introduction

Transportation impacts the lives of all Torontonians. While there is no easy answer to transport people more efficiently in the downtown core there has been an interesting pilot project on one of Toronto's busiest streets, King Street. The King Street pilot project aims to improve transit reliability, speed, and capacity by giving transit priority on King Street from Bathurst Street to Jarvis Street. This capstone project will examine the effects of the King Street Pilot project has made on streetcar transportation in that area. This project will seek to quantify the changes the pilot project has made on travel times, and delays on the 504 streetcar route. Some example questions are: Did the pilot make an improvement to TTC travel time or reduce delays on that route? Data mining techniques using machine learning algorithms like clustering, logistic regression and classification will be used to find meaningful patterns in the dataset.

Literature Review

In recent years, transit systems have been able to collect more quality and higher quantities of information on how transit is used. It is likely that cities will use this information to provide data driven decisions, such as with the City of Toronto to pursue the King St. Pilot Project. The use of Bluetooth sensor tracking, smart card data and modernized Communications and Information Systems (CIS) will provide opportunities for operational efficiencies, improved service and reduced costs.

Many transit systems across the world are using smart cards, which are portable credit card size devices that are used for public transit fare payment. There is much literature of using smart card data to create data models and apply machine learning algorithms to predict transit patterns and usage. For example, by modelling every commuter individually, Nasri Bin Othman at al were able to synthesize highly detailed measurements. This can be used to predict the outcomes of various strategies related to population growth and transportation such as expanded peak hours and crowdedness limits.[1] Lee et al did a case study to improve service and reliability of bus routes using statistical analysis of Singapore's smart card data collected from bus transit trips.[2] Other researchers in Singapore used clustering techniques to exhibit unique and interesting passenger travel patterns for individual train stations. [3] By grouping stations together and quantifying their unique qualities, data mining can help city planners and transit officials implement more strategic and targeted design.

A paper by Kusakabe Takahiko and Yasuo Asakura used the naïve Bayes classifier to help transport operators to monitor and data mine traveler behaviour features observed in smart card data.[4] Another study has proposed a series of efficient and effective data mining approaches with which to model transit riders' travel patterns using smart card data collected in Beijing, China. Some of the algorithms used in this approach were Naïve Bayes Classifier, C4.5 Decision Tree and K-Nearest Neighbor (KNN). [5] An important aspect of the work presented by Aqib, Muhammad, et al is prediction using a distributed computing platform using R and Apache Spark using the London Metro as a case study. This paper proposes a large-scale and fast prediction of metro system characteristics by employing the integration of four leading-edge technologies; big data, deep learning, in-memory computing and GPUs. [6]

A local perspective on transit in Toronto is provided by Chun Kong Yuen, Christopher who produced three models using Ordinary Least Squares regression to investigate transit reliability trends for three major streetcar routes.[7] Another paper using a Canadian dataset used data mining techniques (k means clustering) help to identify and characterize market segments among public transportation users. This study categorized four distinct groups and compared these groups to the different smart card user types, (student, adult and elderly) to compare similarities and differences. [8] This Canadian smart card dataset was used in another research paper using clustering techniques but this time focused more on time of day usage and weekday versus weekend usage. [9]

Lastly, there was an article by Tabassum, Anika Tabassum, et al. that examined data mining possibilities using critical infrastructure systems to build an improved traffic system it is necessary to control or reduce traffic congestion. One example from the article was an algorithm that could be used to detect traffic congestion using the road intersection network, the number of vehicles passing during green lights and a ratio of effective usage of greenlight time.[10]

Dataset

There are three datasets used for this project. Two datasets are from the City of Toronto Open Data (www.toronto.ca/open), more specifically, TTC travel and delay information for the King St. Pilot Project from 2017 – 2018. The third dataset is weather data for Toronto, Ontario provided by the Government of Canada (http://climate.weather.gc.ca). From this dataset the daily temperature and precipitation information from 2017 – 2018 will be used.

The datasets will be restructured and grouped by certain time of day ranges and then joined into one table for exploratory analysis and data modelling. The dataset will focus only on weekday information, thus, excluding weekends and holidays. Our area of interest will be anywhere between King St. and Bathurst St and King St. and Jarvis. We will be looking at 2017 and 2018 only, providing almost a year pre pilot project and post pilot project (the project started Nov 12 2017).

Table 1: Dataset Attribute Descriptions

Attribute	Description
date	Date of measurements
time_of_day	Time of day of measurements: Early, AM Peak, Mid, PM Peak, Evening, Late
day_of_week	Day of the week, eg. Monday, Tuesday, Wednesday.
no_of_screetcars	The number of streetcars on the route during the measurement period
direction	The direction the streetcar was travelling, either Eastbound or Westbound
avg_travel_time	The average travel time of street cars on the route during the measurement period.
no_of_delays	The number of delays the occurred on the route during the measurement period.
avg_delay_time	The average amount of delay time during the measurement period.
avg_speed	The average speed of the streetcars on the route during the measurement period.
temp	The average temperature during the measurement period.
weather_type	The type of weather during that period, either sunny/overcast, rain, snow, or freezing rain etc.
pilot_non_pilot	Label to indicate if the measurement occurred during the pilot or before the pilot project.

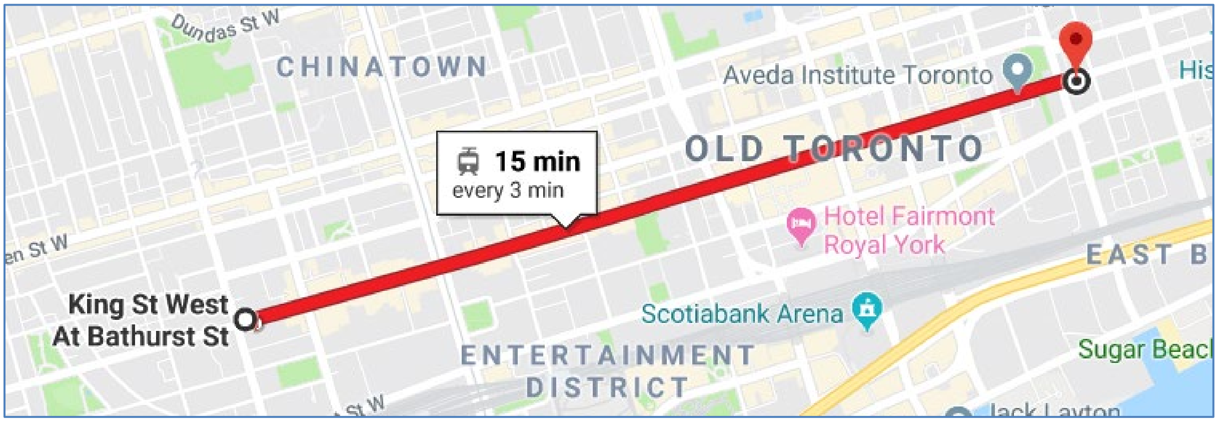


Figure 1: The area of focus for data mining of the King St. Pilot Project

Table 2: Dataset Summary Mean, Median, Min, Max

Attribute	Mean	Min	Max	Median
time_of_day	2	1	5	2
day_of_week	3	1	5	3
no_of_screetcars	15	1	34	16
avg_travel_time	14.6	7.6	50.7	11.2
no_of_delays	1.3	1	5	1
avg_delay_time	18.3	0	1185	8
avg_speed	11.4	3.5	21.7	11.2
temp	8.9	-18.7	28.1	8.1
weather_type	12	2	13	13

Table 3: Correlation Matrix

	temp	avg_travel_ time	avg_ speed	no_of_ streetcars	avg_delay	no_of_ delays	weekday	time_of_day	weather_ type	pilot_ nonpilot
temp	1.000	0.075	-0.038	-0.025	-0.033	-0.057	-0.007	-0.006	0.037	0.097
avg_travel _time	0.075	1.000	-0.849	0.192	0.069	0.122	-0.067	0.019	0.067	0.174
avg_speed	-0.038	-0.849	1.000	-0.403	-0.037	-0.088	0.082	-0.036	-0.071	-0.166
no_of_ streetcars	-0.025	0.192	-0.403	1.000	-0.089	0.030	0.047	-0.271	0.022	0.055
avg_delay	-0.033	0.069	-0.037	-0.089	1.000	0.064	0.031	-0.010	0.037	0.083
no_of_delays	-0.057	0.122	-0.088	0.030	0.064	1.000	-0.018	-0.051	0.047	0.062
weekday	-0.007	-0.067	0.082	0.047	0.031	-0.018	1.000	-0.014	0.006	0.018
time_of_day	-0.006	0.019	-0.036	-0.271	-0.010	-0.051	-0.014	1.000	0.013	-0.024
weather_type	0.037	0.067	-0.071	0.022	0.037	0.047	0.006	0.013	1.000	-0.029
pilot_nonpilot	0.097	0.174	-0.166	0.055	0.083	0.062	0.018	-0.024	-0.029	1.000

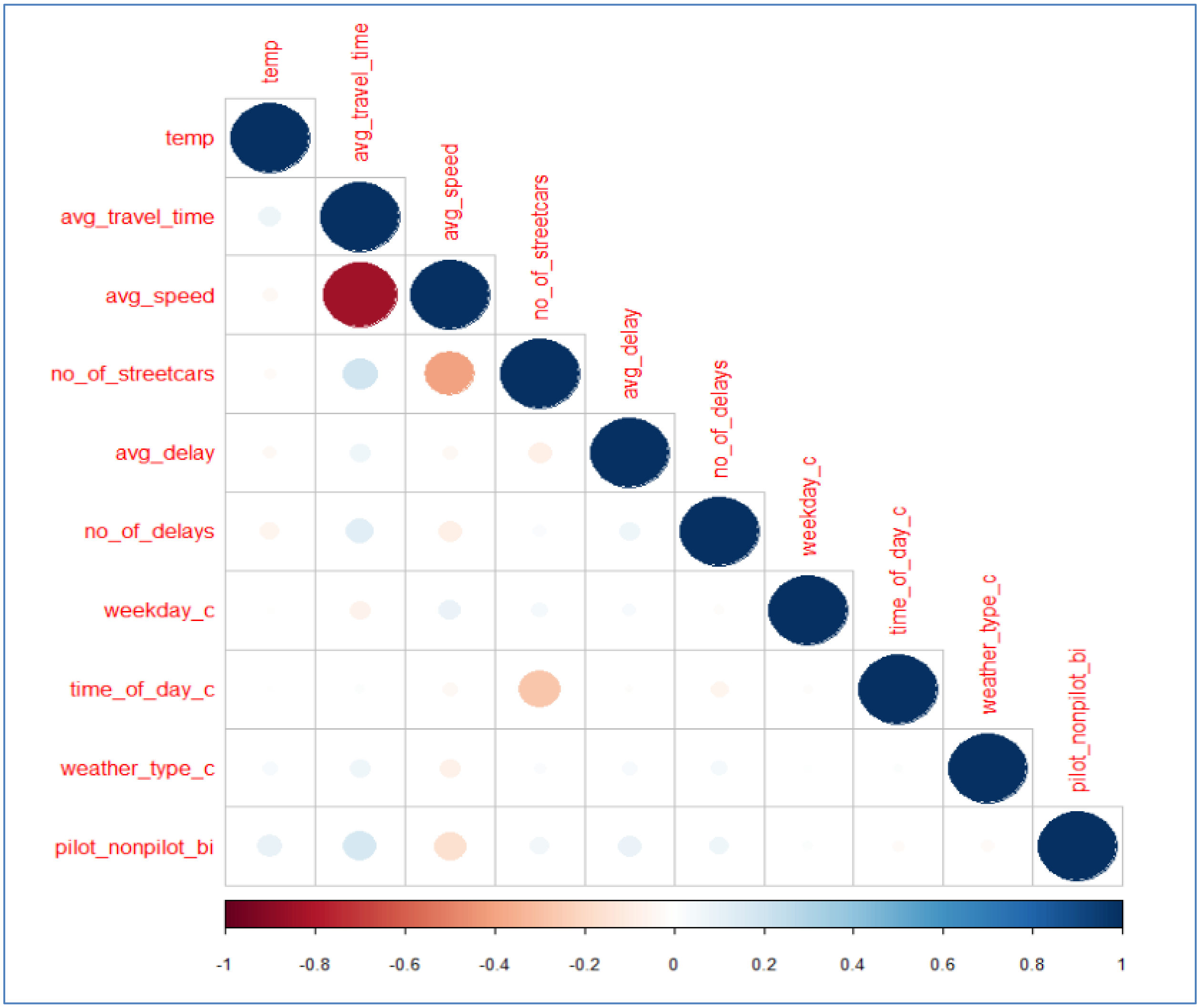


Figure 2: Visualization of Correlation Matrix

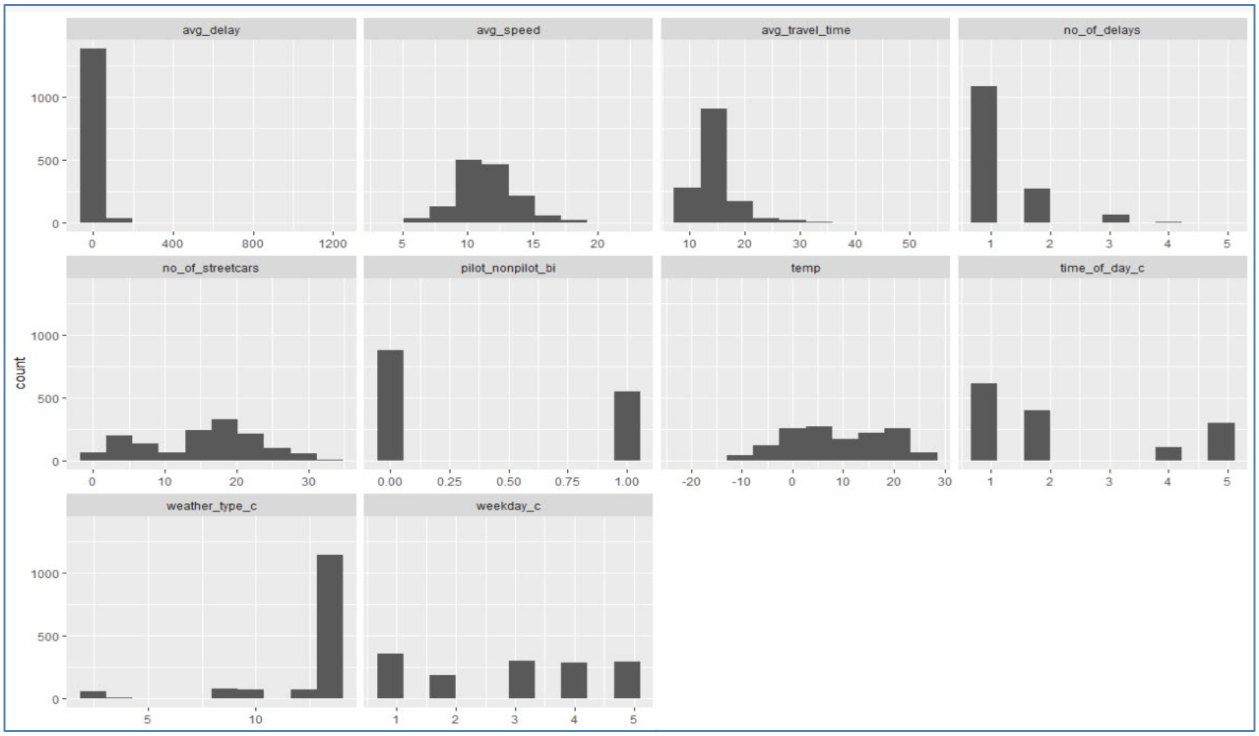


Figure 3: Histograms of dataset attributes

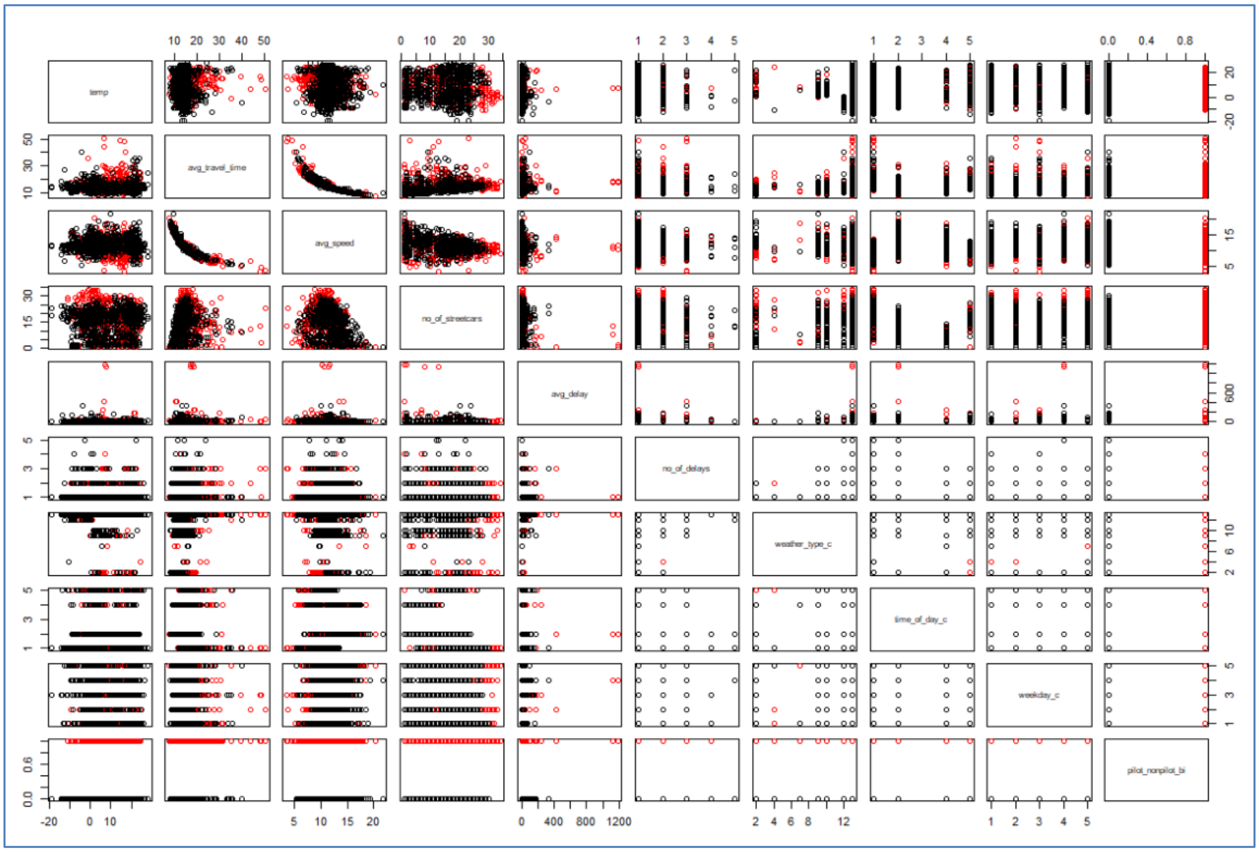


Figure 4: Visualization of dataset with pilot and non-pilot represented as red and black respectively.

Approach

This project's approach starts with importing, cleaning and restructuring the data. Data cleaning includes removing missing values, standardizing naming conventions and grouping measurements based on certain times of the day. Then, exploratory analysis is performed to become familiar with the dataset, identify outliers and relationships between attributes. Then data modelling will be used and then the results will be interpreted into meaningful observations.

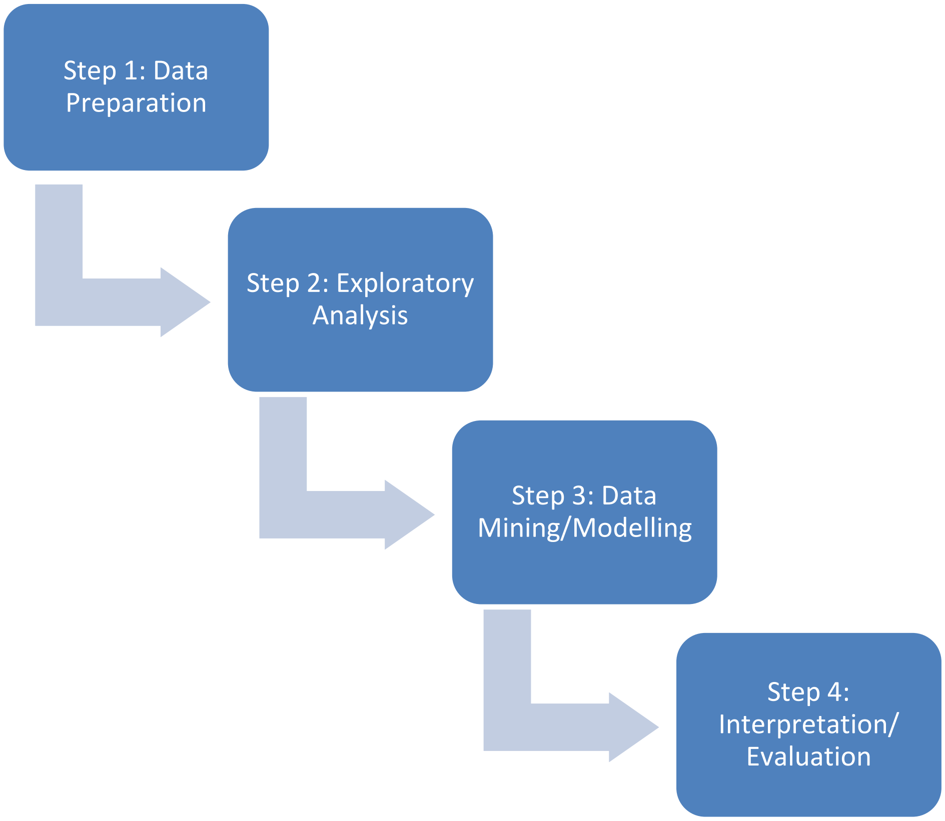

Step 1: Data Preparation
Step 2: Exploratory Analysis
Step 3: Data Mining/Modelling
Step 4: Interpretation/ Evaluation

Figure 5: Step by step approach for project.

Step 1: Data Preparation

The first step is to collect the data sources from the City of Toronto and the Government of Canada websites for both 2017 and 2018. The data is imported into the software tools. Firstly, missing and NA values will be removed or estimated as necessary. In the case of the transit delay data, only delays that occurred with route 504 between Bathurst St. and Jarvis St. will be extracted for use in the dataset. Unnecessary attributes will be removed from the datasets, for example in the transit delay dataset attributes such as vehicle number and type of delay will be removed. In the case of the weather data, only the temperature and type of weather will be extracted for use in the main dataset. And for the TTC travel time dataset, attributes such as vehicle number, run number, and route number will be removed. Then the measurements will be restructured and grouped base on time of day, for example, a measurement taken at 5AM will be grouped into a category called "EARLY". The datasets will then be joined based on date and time of day into the main dataset to be used for exploratory analysis and data mining.

Step 2: Exploratory Analysis

The step will help to better understand the data and to discover initial patterns in addition to getting familiar with the dataset. During this step, values will be plotted for visualization using scatter plots and box plots. This will also help in identifying outliers. Another part of this step will be to use a simple spearman correlation to uncover any potential relationships between attributes. Anything of interest will be included in the final report.

Step 3: Data Mining/Modelling

In this step, machine learning K nearest neighbor clustering technique will be used to identify patterns in the data. During this step, measurements will be clustered into groups in order to identify unique characteristics. Logistic regression will be used to identify relationships between attributes to identify which attribute had the most impact on streetcar travel time.

Step 4: Interpretation/Evaluation

The final step will evaluate the accuracy of the models and interpret what the correlations and relationships between attributes mean to the impact of the King St pilot project. Can this analysis show improvements in transit due to the pilot project?

Results

Model Selection

The dependent variable chosen for these models was a label of “pilot” or “non pilot” to predict if a measurement was from during the pilot period or before the pilot period. This dependent variable was converted into binary with 0 = “pilot” and 1 = “non pilot”. Two classifier models were chosen in order to predict the dependent variable, the knn classifier and the binomial logistic regression classifier. These models are easy to implement, easily interpretable, used widely by data analyst and scientist. The classifiers can also work with categorical information which was present in the dataset. These classifiers are simple and effective and have feature independence assumptions.

Model parameters

Tests were used in order to optimize the model parameters in order to increase accuracy. Precision will be prioritized over recall so that the model can have a high quality of true positives. The error rate was used for the KNN model to find the optimum number of k centers to produce the lowest error rate which in this case is 0.28. The results of the k centers error rate test can be seen in figure 6. The figure shows that as the number of centers is increased above 5 the error rate increases. This model performed well with an accuracy of 71%.

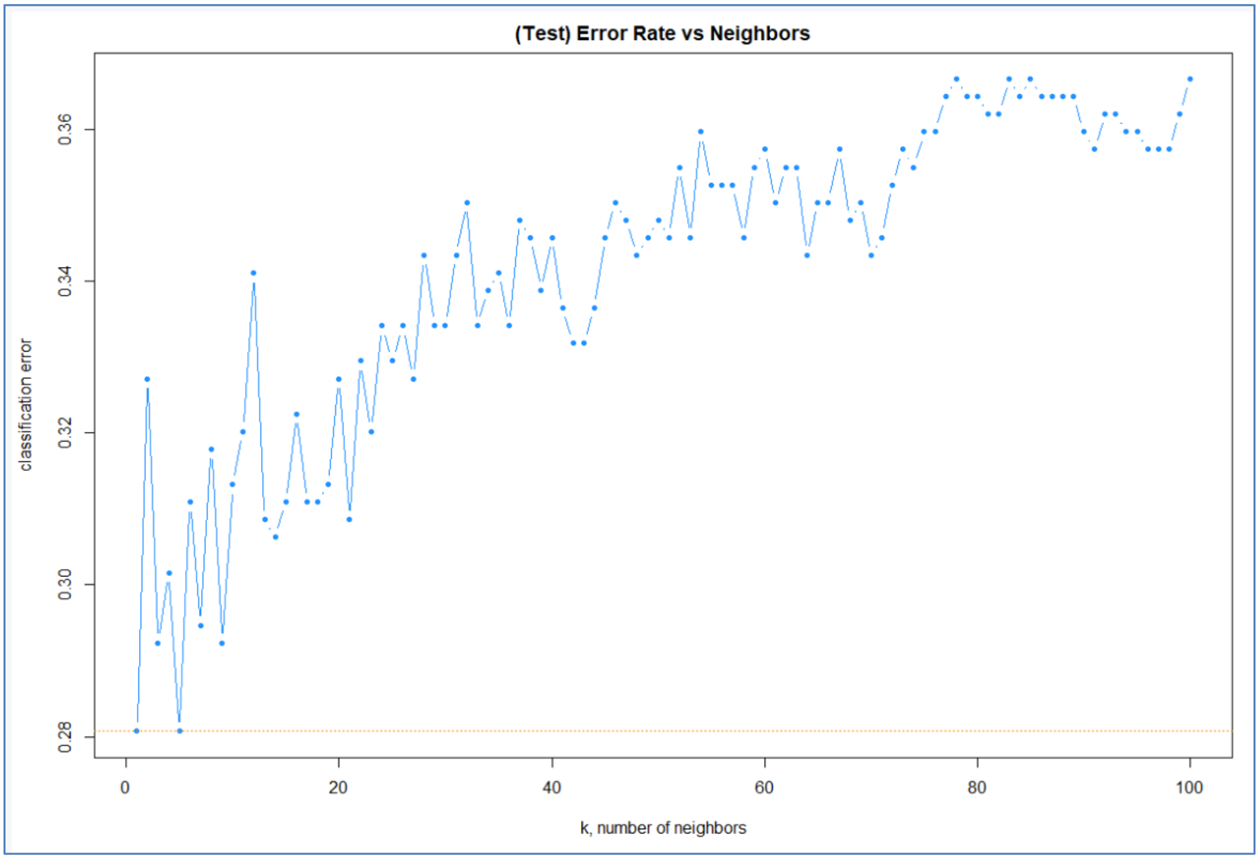


Figure 6: Error rate of KNN model versus the number of k centers.

For the binomial logistic regression model the probabilities from the model can be converted to predictions using a threshold value. The threshold value was determined by using the receiving operating characteristic (ROC) curve. This curve can be seen in figure 7, with the x-axis as the False Positive Rate and the y-axis as the True Positive Rate. The threshold used was selected in order to maximize the True Positive Rate while keeping the False Positive Rate low. The threshold used was 0.4, as this provided the highest True Positive Rate and the lowest False Positive Rate.

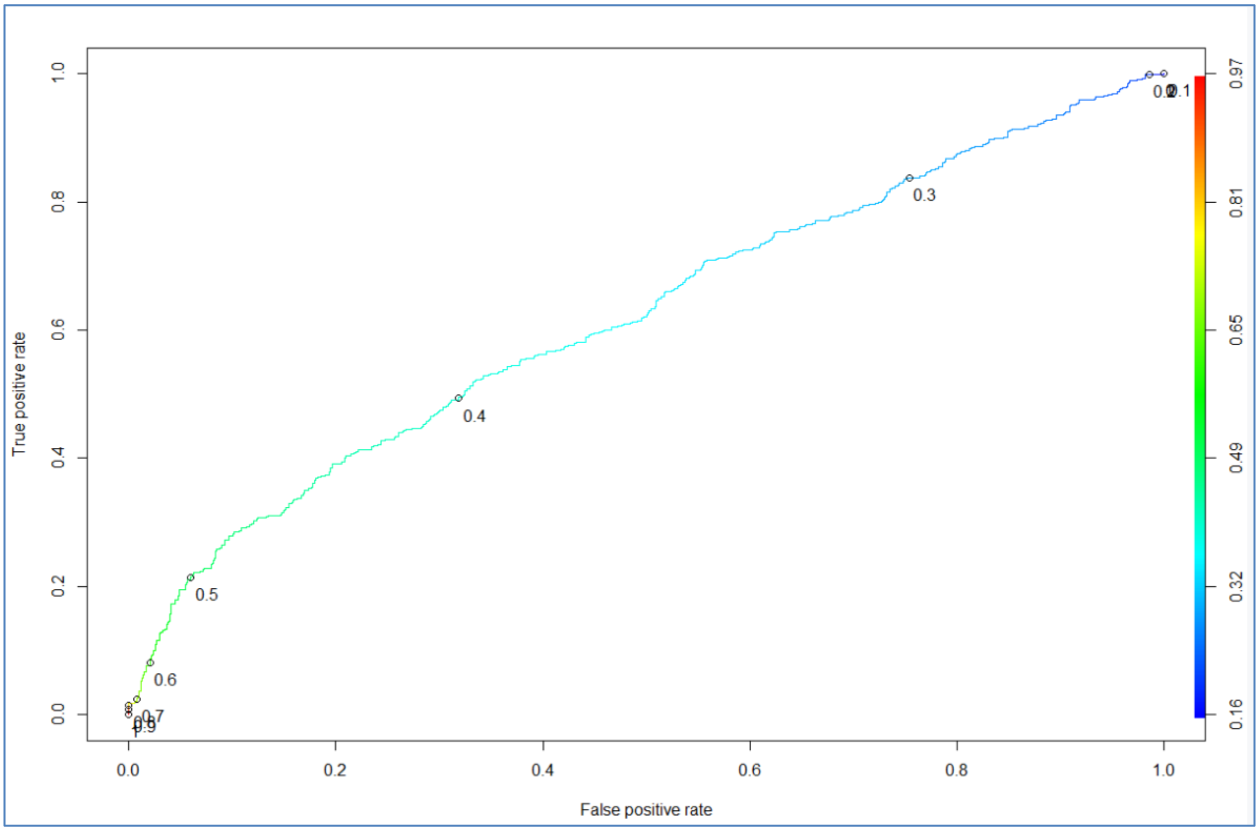


Figure 7: True positive versus false positive rate

Model Evaluation

KNN

This model had an accuracy of 71% and the precision value was higher than the recall value. Backwards elimination was used to select only relevant features. These features were travel time, speed, number of streetcars, delay time, number of delays and weather type. These features provided the optimum subset of information which added value to the model. Since accuracy does is not always the best metric to use when evaluating a model, the F1 Score was also calculated. This model had an F1 Score of 60% which is in an acceptable range but is not very high.

		Predicted	
		0	1
Actual	0	213	46
	1	79	93

Table 4: Confusion Matrix for KNN Model

Statistic	Value (%)
Accuracy	71
Precision	67
Recall	54
F1-Score	60

Table 5: Evaluation Metrics for KNN Model

Binomial Logistic Regression

The logistic regression model did not perform as well as the KNN model with an accuracy of only 66%. Looking at the F1 Score of this model, it performs more poorly at only 14%. This model was more successful at classifying true negatives than it was at classifying true positives and it mislabeled many datapoints as false when they were true. Therefore, the model did not perform very well.

		Predicted	
		0	1
Actual	0	829	53
	1	435	118

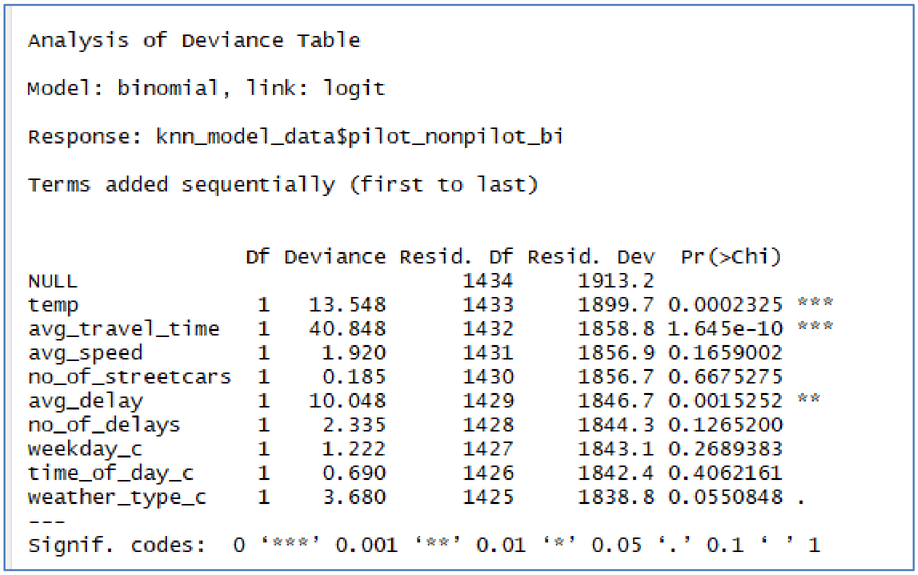
Table 6: Confusion Matrix for Logistic Regression Model

Statistic	Value (%)
Accuracy	66
Precision	48
Recall	21
F1-Score	14

Table 7: Evaluation Metrics for Logistic Regression Model

The receiving operating characteristic (ROC) in figure 6 is used to measure the logistic regression model performance in addition to accuracy, precision, recall and F1 Score. The ROC is used to calculate the Area Under the Curve (AUC) which will provide an evaluation metric. The AUCROC is 0.66 which indicates a close to worthless classifier (0.5 being the lowest acceptable threshold). This reinforces that this model was not effective in classifying the pilot vs. non-pilot data. This could potentially be due to the large amount of categorical information in the dataset. Or it could be because there were too many independent variables.

Although the logistic regression model was unsuccessful at accurately classifying the dataset, it did provide some insights to the relationships between the independent and the dependent variable. Figure 8 displays the results of the logistic regression model in the R software. The p-values which are used for validating a hypothesis help to describe which independent variables are significant or insignificant to predicting the dependent variable. The smaller the p-value the more significant the variable. The hypothesis in this model was if the independent variable had an impact on the data before and after the pilot project which is the dependent variable. The results of the logistic regression model indicate that the temperature, average travel time and the average delay time are statistically significant in predicting the pilot or non-pilot value. The temperature variable only indicates that the weather in the year before the pilot was different than the weather in the year after the pilot. This is to be expected as weather does change year over year. What is most interesting is that this model does show that the average TTC travel time and average TTC delay time was in fact impacted by the pilot project. This is what was intended by the pilot project and can thus point to it being a success.



```
Analysis of Deviance Table

Model: binomial, link: logit

Response: knn_model_data$pilot_nonpilot_bi

Terms added sequentially (first to last)


                  Df Deviance Resid. Df Resid. Dev  Pr(>Chi)    
NULL                               1434     1913.2              
temp               1   13.548      1433     1899.7 0.0002325 ***
avg_travel_time    1   40.848      1432     1858.8 1.645e-10 ***
avg_speed          1    1.920      1431     1856.9 0.1659002    
no_of_streetcars   1    0.185      1430     1856.7 0.6675275    
avg_delay          1   10.048      1429     1846.7 0.0015252 ** 
no_of_delays       1    2.335      1428     1844.3 0.1265200    
weekday_c          1    1.222      1427     1843.1 0.2689383    
time_of_day_c      1    0.690      1426     1842.4 0.4062161    
weather_type_c     1    3.680      1425     1838.8 0.0550848 .  
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Figure 8: Output of logistic regression model in R including P-values.

Conclusions

This project focused on data collected from the King St transit pilot that occurred in Toronto, Ontario. Three datasets were obtained from open sources and were joined into a single tabular structure and subjected to data preparation for exploratory analysis. Predictive models were used to classify data points from before the pilot and after in order to identify key features of the dataset. The two models used were the KNN and binomial logistic regression model. These models were evaluated using accuracy, recall, precision and AUROC. The KNN performed better than the logistic regression in how accurate the data was classified. However, the logistic regression uncovered which features had the lowest p value thus the highest significant difference between pilot and non-pilot data. This could indicate an impact caused by the pilot project on TTC travel time and TTC delay time.

References

[1] Nasri Bin Othman, et al. “Simulating Congestion Dynamics of Train Rapid Transit Using Smart Card Data.” A*STAR Computational Resource Centre, Institute of High Performance Computing, Agency for Science, Technology and Research, 17 Feb. 2014, arxiv.org/pdf/1402.3892.pdf.

[2] Lee, D.H., Sun, L., Erath, A. (2012) Study of Bus Service Reliability in Singapore Using Fare Card Data. Proceedings of 12th Asia Pacific ITS Forum & Exhibition, Kuala Lumpur. https://www.research-collection.ethz.ch/handle/20.500.11850/53911.

[3] LEE, Roy Ka Wei, and Tin Seong KAM. “Time-Series Data Mining in Transportation: A Case Study on Singapore Public Train Commuter Travel Patterns.” Research Collection School Of Information Systems, Singapore Management University, Oct. 2014, ink.library.smu.edu.sg/cgi/viewcontent.cgi?article=3447&context=sis_research.

[4] Kusakabe Takahiko, and Yasuo Asakura. “Behavioural Data Mining of Transit Smart Card Data: A Data Fusion Approach.” Elsevier Ltd., Department of Civil Engineering, Tokyo Institute of Technology, Japan, 22 May 2014, journals-scholarsportal-info.ezproxy.lib.ryerson.ca/pdf/0968090x/v46icomplete/179_bdmotscdadfa.xml.

[5] Ma, Xiaolei, et al. “Mining Smart Card Data for Transit Riders’ Travel Patterns.” Transportation Research, Elsevier Ltd., 18 July 2013, journals-scholarsportal-info.ezproxy.lib.ryerson.ca/pdf/0968090x/v36icomplete/1_mscdftrtp.xml.

[6] Aqib, Muhammad, et al. “Rapid Transit Systems: Smarter Urban Planning Using Big Data, In-Memory Computing, Deep Learning, and GPUs.” MDPI, Multidisciplinary Digital Publishing Institute, 14 May 2019, www.mdpi.com/2071-1050/11/10/2736.

[7] Chun Kong Yuen, Christopher. “Exploring Transit Performance And Traffic Congestion in ...” Ryerson University Library, Ryerson University, Jan. 2017, digital.library.ryerson.ca/islandora/object/RULA:6388.

[8] Agard, Bruno, et al. “MINING PUBLIC TRANSPORT USER BEHAVIOUR FROM SMART CARD DATA.” *IFAC Proceedings Volumes*, Elsevier, 21 Apr. 2016, www.sciencedirect.com/science/article/abs/pii/S1474667015359310.

[9] Morency, Catherine, et al. “Measuring Transit Use Variability with Smart-Card Data.” Transport Policy, Pergamon, 12 Mar. 2007, www.sciencedirect.com/science/article/pii/S0967070X07000030.

[10] Tabassum, Anika Tabassum, et al. “Data Mining Critical Infrastructure Systems: Models and Tools.” IEEE Intelligent Informatics, IEEE Intelligent Informatics Bulletin, Dec. 2018, people.cs.vt.edu/~badityap/papers/cis-ieeeiib18.pdf.