Predictive Analytics for Demand Responsive Para-transportation

Yaozhong Huang, Yihe Chen, Kehan Yu, Junting Chen

Team RTS, DSCC 383W

Goergen Institute for Data Science

University of Rochester

1. Introduction

1.1 Background

People nowadays live in an era that is highly digitalized and information-based. Their living standard and quality are highly dependent on the effectiveness of data and information associated. Also, in the view of business, companies can increase their revenues by building optimization models based on their customer data. In the era of data overload, both Individuals and organizations want to benefit from leveraging the power of data[1].

As a considerable part of people's daily routine and city operation, the field of public transportation is recently exploring its new form in the age of information. According to contemporary transportation definitions, there are two major types of public transportation systems: conventional public transport systems and para-transportation. Conventional public transportation is typically managed on a fixed schedule, operated on established routes, and charged a posted fee for each trip. In contrast, the paratransit service includes all systems of urban transport, which supplement conventional public transit by providing individualized rides without fixed routes or timetables[2]. Specifically, para-transportation provides service for people with disabilities that have a functional impairment with using the fixed-route service some or all of the time. Additionally, in a more general scenario, it includes services such as carpools, specialized commuter bus services on a subscription basis, as well as taxi cab related services.

The availability of increased computational power, analytical approaches, and the collection of massive amounts of data have redefined the value of the machine learning-based methods[3] for addressing the emerging problems and improving the practical optimization issues. Empowered by the data science techniques, this project aims to apply modern machine learning techniques to create a productive schedule for Demand Responsive Para-transportation by predicting the customer's cancellation based on their reservation information and some external information such as weather data. With the acknowledgment of past work in the machine learning area, those machine learning solutions have already begun their promising marks in the transportation industry, generating a more precise, intuitive, and reliable output than traditional models. The underlying goals for these applications of machine learning techniques are to reduce industry operation cost, improve service effectiveness, diminish human errors, and meanwhile mitigate unfavorable environmental impacts[4].

Our sponsor RTS, the Rochester-Genesee Regional Transportation Authority, is a New York State public-benefit corporation that provides transportation services in the eight-county area in and around Rochester. RTS Access also provides complementary paratransit service for people with disabilities, which is a demand-responsive, shared-ride public transportation service.

1.2 Project Goal

In this project, we focused on RTS's para-transportation service and specifically the cancellation of the booked trips. In particular, we are going to discover the relationships between customer reservation cancellation rate and several related attributes, such as "trip purpose", "trip type", "reservation time", "seasonal pattern", and so forth. Besides the exploratory data analysis, we also tried different machine learning models to decide the best-fit one for predicting the possible cancellation of the upcoming trips. The input of our model includes features such as reservation information, customer information, trip information, and external factors. In this report, we are going to present a process of progressive implementation and improvement of the models, and finally, determine recommendations for mitigation strategies.

2. Data Set

2.1 Dataset description

Two datasets were used in this project. One was the operational dataset from our sponsor RTS, which contains a total of 102754 observations, 21 explanatory variables from May 17th, 2021 to December 5th, 2021. The other was the weather dataset acquired from NOAA (National Oceanic and Atmospheric Administration), including weather data of the Rochester area in the corresponding time span.

2.1.1 Internal data

This part of data was acquired from our sponsor RTS, including 21 descriptive features. Table. 1 shows the variables contained and their corresponding brief description.

Table. 1 Data Definitions of Internal Data

Variable Name	Definition	Variable Name	Definition
Run Name	The run number for the trip	PassOn	The number of customers who boarded the bus
EV Order	Order of Events for the Run	SpaceOn	Boarding requirements for the customers
Client ID	Customer ID Number	Distance	Direct distance
Travel Date	Date of travel	Sched Status	Trip status
Cre_D	Date the trip was requested	Pick City	Customer pickup address

Mod_D	Date the record was modified	Pick Zip	Pickup zip code
Mod_T	Time the record was modified	Pick Poly	Pickup polygon area
Req_T	Customer requested time	Drop City	Customer drop-off address
Neg_T	Customer negotiated time	Drop Zip	Drop-off zip code
Subtype	Trip type	Purpose	Trip purpose
Drop Poly	Drop-off polygon area		

2.1.2 External data:

- The first column is the date. We used this column as the index to merge the weather data with the internal dataset.
- Columns 2 to 7 are the temperature information of each day, including the maximum temperature, minimum temperature, average temperature, and temperature departure, which is the difference between the highest and lowest temperature in a day.
- Column 8 is the precipitation of a day.
- The last two columns are the snow information of a day. New snow denotes the snowfall of the day, and snow depth is the cumulative depth of snow on that day.

Table. 2 Weather Data

	Date	Max_temp	Min_temp	Avg_temp	temp_departure	HDD	CDD	Precipitation	New_snow	Snow_depth
0	2021-05-01	59	33	46.0	-7.5	19	0	0.00	0.0	0
1	2021-05-02	62	50	56.0	2.1	9	0	0.03	0.0	0
2	2021-05-03	68	45	56.5	2.2	8	0	0.03	0.0	0
3	2021-05-04	71	45	58.0	3.3	7	0	0.07	0.0	0
4	2021-05-05	57	47	52.0	-3.1	13	0	0.00	0.0	0

2.2 Data Cleaning

In the RTS dataset, 4 columns have missing values: Pass On (the number of customers who boarded the bus), Space On (boarding requirements for the customers), Pick Zip (Pickup zip code), and Drop Zip (Drop-off zip code). To deal with missing values, we deleted 1053 rows, which are about 1% of the whole dataset, remaining 102809 rows. Besides fixing the missing values, we also removed the units from the dataset, such as we dropped “mi” in the distance variable.

In the weather dataset, some variables contain a value of “T” or “M”, which stands for “trace” or “missing”. For simplicity, we replaced these values with 0.

2.3 Feature engineering

2.3.1 Creating a label for the cancellation

Since our project is aiming at predicting the cancellation of scheduled trips, it is important to label each record as either canceled or not. This information is extracted from the Sched Status column, which includes 18 different types of trip statuses such as regularly performed, performed but not show up, canceled, etc. We assigned 1 to all canceled trips, and 0 to all performed trips.

2.3.2 Grouping request time into hours

Request time is the scheduled time of the trip. We transformed the request time from seconds to hours as a more informative form for the later Exploratory Data Analysis session.

2.3.3 Grouping clients by their previous cancelation rate

Clients with high past cancellation rates are more likely to cancel again than those with low cancellation rates. We assigned the clients into three groups (low, medium, and high cancelation rate group) based on their past cancelation rate. We use kernel density estimation to cluster the clients, which is good for 1-dimensional data. As shown in Fig.1, the kernel density estimation suggests that the low cancellation group has a cancellation rate lower than 0.4, the medium cancellation group has a rate between 0.4 and 0.9, and the high cancellation group has a rate higher than 0.9.

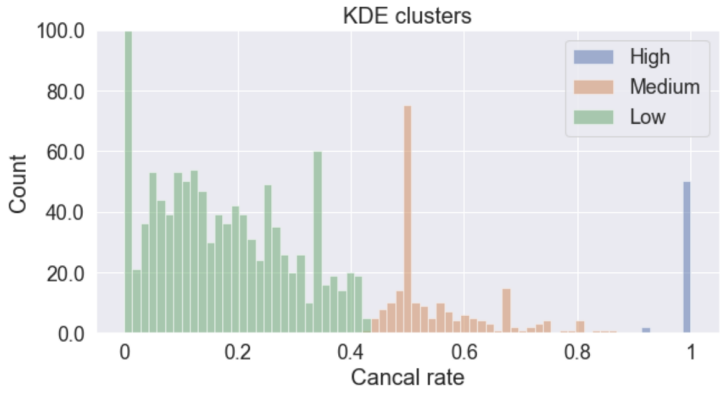


Fig.1 Client segmentation result using Kernel Density Estimation

2.3.4 Transforming Datetime object to informative variables

We have 3 columns of Datetime variables: Travel Date (Date of travel), Creation Date (Date the trip was requested), and Modified Date (Date the record was modified). To perform exploratory data analysis on these variables, we transformed each column into three integer-type columns, indicating the year, month, and day of each date. After doing so, we also created a new column called weekend to indicate if that date is a weekend.

2.3.5 Encoding trip type and purpose

We performed one-hot encoding for trip type and purpose, creating dummy variables for each type and purpose.

2.3.6 Aggregating clients

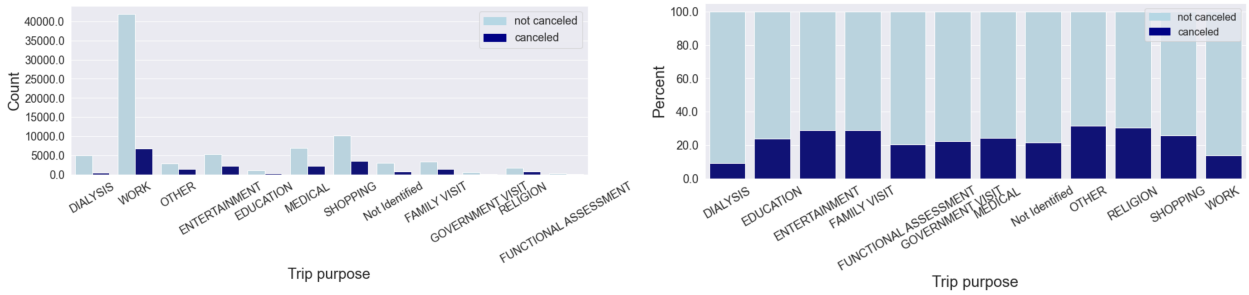
There are four types of passengers: Client, Fare-paying guest, Child or under 5, and Personal Care Assistant. We aggregated all trips by the need for lifts, the need for personal care assistants, the number of children, and the total number of passengers, and created four columns to store this information.

3. Exploratory Data Analysis

In order to have a better understanding of how different features relate to cancellation, in this section, we will analyze from eight aspects, including the trip purpose, the type of trip, the number of days the trip was booked in advance, the time of day and the day of the week the trip was scheduled, and external weather conditions such as temperature, snowfall, and precipitation.

3.1 Trip Purpose

The amount and percentage of canceled and uncanceled trips for 12 trip purposes are shown in Fig.2 (a) and (b), respectively. From Fig.2 (a), we found that most trips were booked for work. When looking at the percentage of cancellations, trips booked for dialysis and work have the lowest cancellation rate, while trips with entertainment, family visit, and religious purposes were more likely to be canceled.

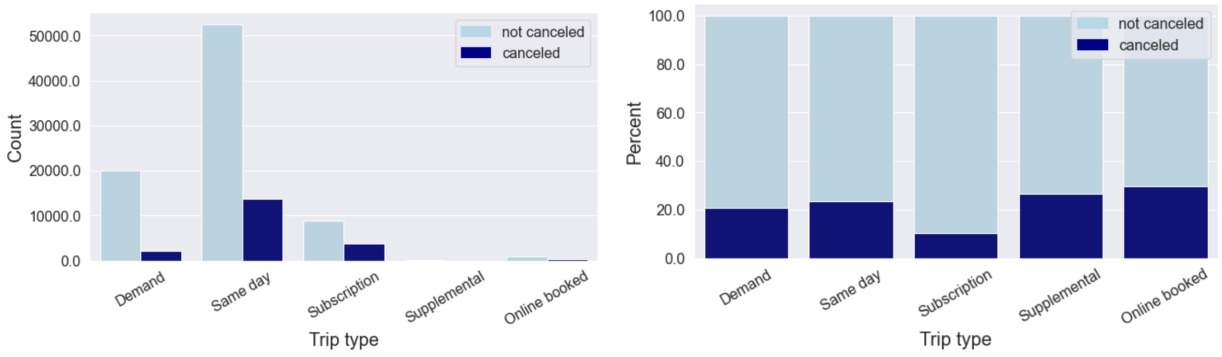


(a) (b)

Fig. 2 Canceled and not canceled trips for different trip purposes
(a) Side-to-side Barplot showing the count of canceled versus not canceled trips for each feature value
(b) Stacked Barplot showing the percentage of canceled versus not canceled trips

3.2 Trip Type

We found that the same-day trips made up a large part of the total reservations based on fig.3 (a). On the other hand, fig.3 (b) shows that subscription trips have the lowest cancellation rate at around 8%, whereas trips booked online have the highest cancellation rate at around 30%.

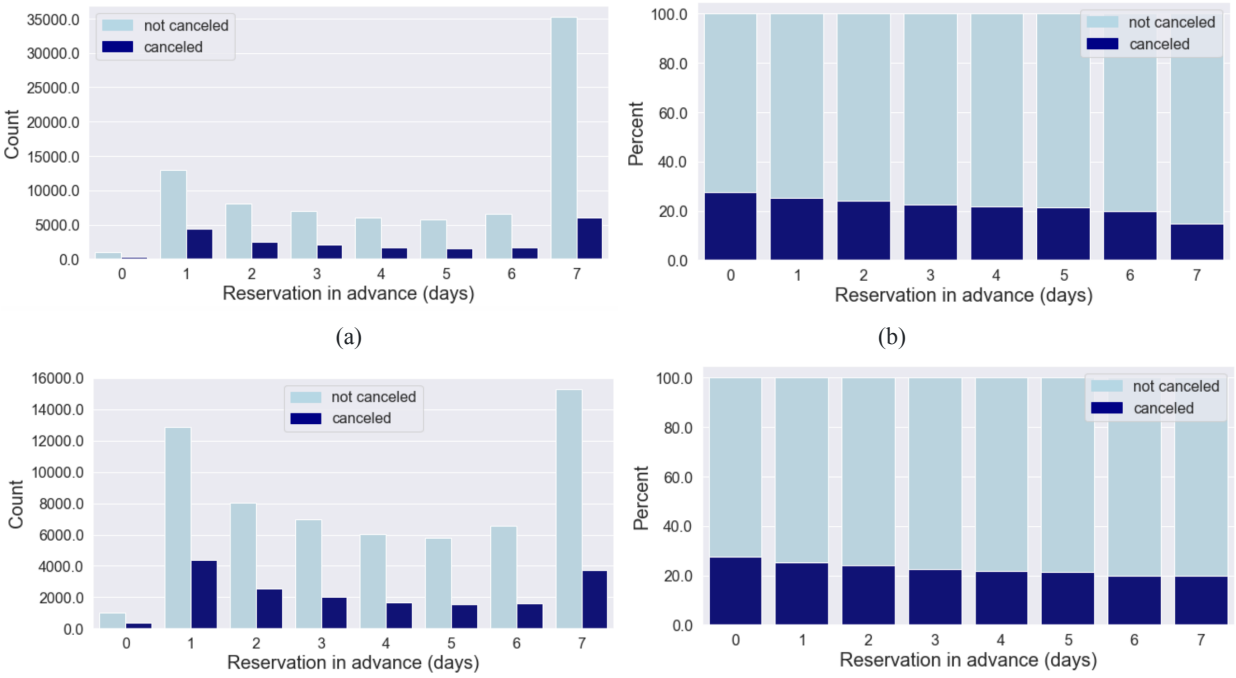


(a) (b)

Fig. 3 Canceled and not canceled trips for different trip types
(a) Side-to-side Barplot showing the count of canceled versus not canceled trips for each feature value
(b) Stacked Barplot showing the percentage of canceled versus not canceled trips

3.3 Reservation in Advance

Based on Fig.4 (a) and (b), most trips were scheduled 7 days in advance, which includes a large proportion of subscription trips with lower cancel rates as the analysis of trip type showed. After removing subscription trips as Fig.4 (c) and (d) show, there was no obviously lower cancel rate for trips reserved 7 days in advance. Therefore, subscription trips might be an important factor for predicting cancellation.



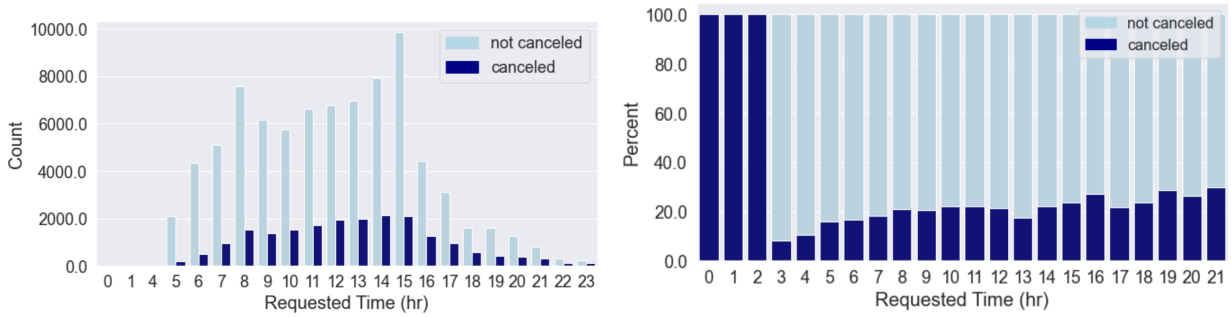
(a) (b)

(c) (d)

Fig. 4 Canceled and not canceled trips for different number of days the trip was scheduled in advance
(a) Side-to-side Barplot showing the count of canceled versus not canceled trips for each feature value
(b) Stacked Barplot showing the percentage of canceled versus not canceled trips

3.4 Requested Time

Most requests are from 8 am to 3 pm as shown in Fig. 5 (a). Also, there was an increasing trend in cancel rate throughout the day, reaching a peak from midnight to 2 am with a 100% cancellation rate. On the other hand, the sample size for trip reservations during midnight to 2 am was very small. Therefore, the cancellation rate may vary significantly as more data is added to the analysis.



(a) (b)

Fig. 5 Canceled and not canceled trips for different requested time in hours
(a) Side-to-side Barplot showing the count of canceled versus not canceled trips for each feature value
(b) Stacked Barplot showing the percentage of canceled versus not canceled trips

3.5 Requested Day

In order to determine any seasonal pattern in cancellation, we looked at the daily cancel rate from May to December. We observed a 7-day seasonal period from Fig. 6. In other words, there was a weekly pattern with the highest cancellation rate on Sundays. As the day in a week can affect the cancel rate, we used weekday as one of the model input features.

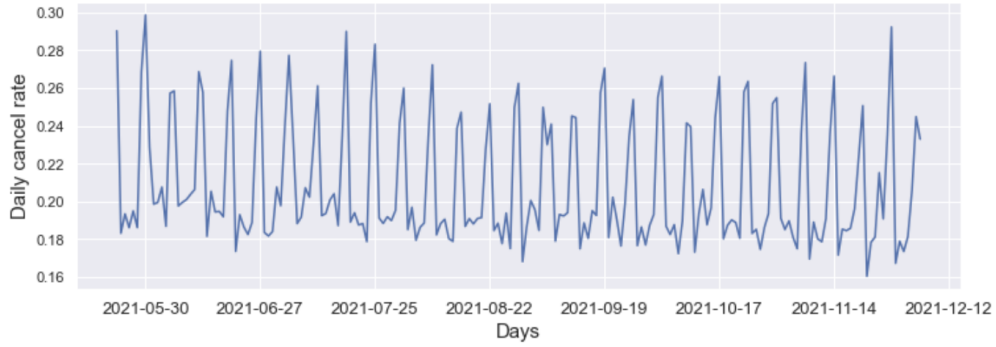
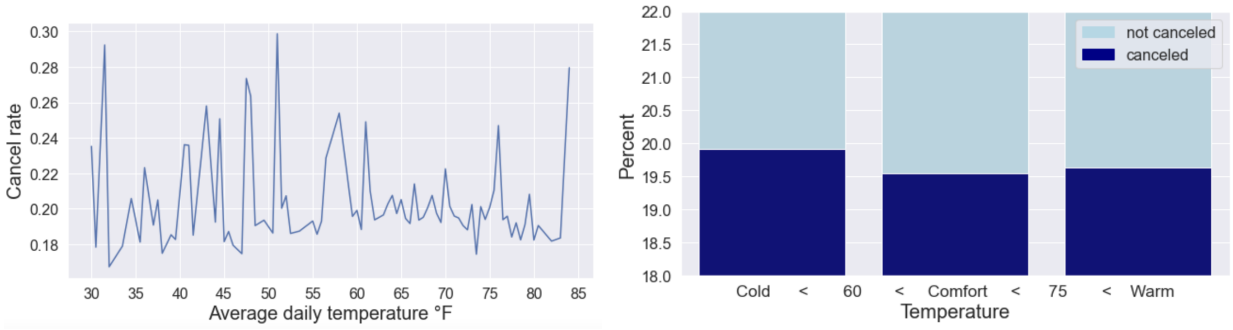


Fig. 6 Daily cancel rate for days from May to December

3.6 Average Temperature

As mentioned in the previous section, we were interested in finding whether weather could play a role in the cancellation. However, it is hard to determine any meaningful pattern between cancel rate and average daily temperature with our current data. According to Fig. 7 (a), the cancellation rate fluctuates when temperature changes, and there was no significant difference in cancellation rate among the three temperature groups as shown by Fig. 7 (b). Relatively speaking, trips scheduled on comfort-temperature days have the lowest percent canceled, but the difference between the comfort and cold-temperature group was still only within 1%.



(a) (b)

Fig. 7 Temperature versus cancelation
(a) Cancel rate on days of different temperatures
(b) Stacked Barplot showing the percentage of canceled versus not canceled trips for different temperature groups. Group Cold contains temperature lower than 60 degree Fahrenheit; Group Comfort contains temperature from 60 to 75 degree Fahrenheit; Group Warm contains temperature higher than 75 degree Fahrenheit

3.7 Snowfall

Based on six days of recorded snow data in November and December, the cancellation rate seems to be higher when there's more snowfall (Fig. 8), but we cannot provide any conclusive finding due to limited data. More data is needed for further analysis.

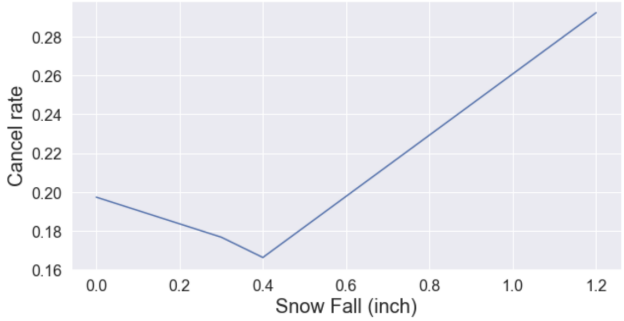
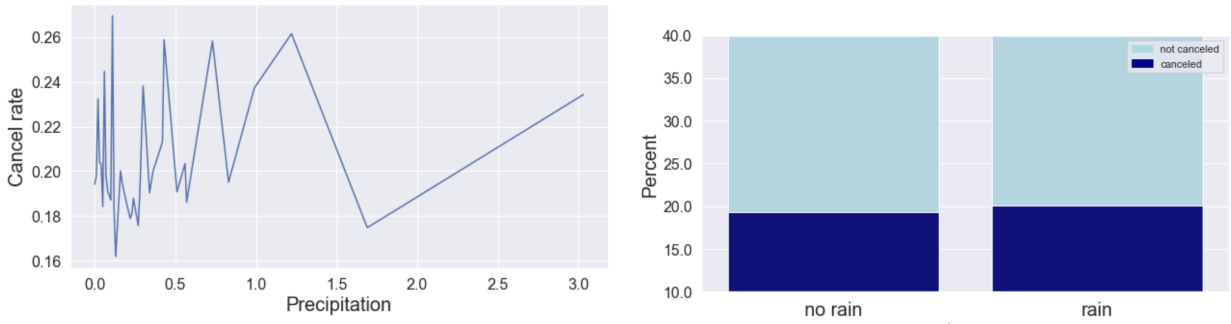


Fig. 8 Snowfall versus cancelation rate

3.8 Precipitation and Rainfall

There is no obvious pattern in cancellation observed when precipitation changes in Fig. 9 (a). Fig. 9 (b) illustrates that the percent of trips canceled on rainy days also wasn't significantly different from no rain days.



(a) (b)

Fig. 9 Precipitation and rain versus cancelation
(a) Cancel rate on days of different precipitation
(b) Stacked Barplot showing the percentage of canceled versus not canceled trips for no rain group and rain group

3.9 Summary

In this section, we further explored and analyzed the cancellation pattern with different variables and obtained valuable information. First, trips booked for work and dialysis were the least likely to be canceled. Moreover, regular subscription trips that were automatically reserved 7 days in advance were found to have the lowest cancellation rate among five trip types. In contrast, trips booked on the same day as the travel time are the most likely to be canceled by customers, and the cancellation rate gradually decreases as trips were booked more days in advance, with the lowest cancellation rate landing on the 7th day even after taking subscription trips out of the equation. As for the time of the day, 8 am to 3 pm seems to be a window in which most of the trips were scheduled, and the cancellation rate starts to increase from 3 am and peaks around midnight. As for the time of the week, Sundays generally have the highest cancellation rate.

In the attempt to analyze cancellation with external weather data, we were not able to conclude any significant patterns. There seems to be a positive correlation between snowfall and cancellation, but more data is needed to back up the initial findings. The time of the year can not only affect weather, but it also has a potential influence on many other variables such as trip purpose and trip type. For more in-depth time series analysis, we need at least one year's worth of data to obtain a full cycle, then we can begin to account for the variation from seasonality, summarize the yearly pattern, and incorporate it into our model.

4. Model development

The objective of this project was to predict whether a trip would be canceled or not, given the data provided by RTS. Therefore, we chose the supervised learning techniques, the classification models, to predict the binary cancellation outcome.

4.1 Modeling Setup

Before reaching the modeling part, we first finished several setups for later analysis.

4.1.1 Train-Test Set Splitting

Following the general process of training a classification model, we first split our dataset into two parts: training and test sets. The training set contains 80% of the data that is used for training models. The test set contains 20% of the data to test the trained models.

4.1.2 Input of the model

To help better understand the model, we grouped the inputs of the prediction models into four categories (shown in Table.3): client information, trip information, reservation information, and weather data.

Table. 3 Model Input Features

Client Information - Client ID: customer ID number - Trip Type: dummy variables created for trip type - Trip Purpose: dummy variables created for the trip purpose - PCA: whether the customer needs a personal care assistant - Lift: whether the customer needs a lift on that trip - Child: number of children on that trip - Total number: number of passengers on that trip	**Trip Information** - Distance: direct distance of the trip - Weekend: whether the trip is on the weekend or not - Time: the trip time of each trip
Reservation Information - Difference: time difference between reservation date and actual perform date	**Weather Data** - Temperature - Precipitation - Snow

4.1.3 Evaluation Metric

To estimate the model performance, we have to choose our evaluation metric. In this project, our evaluation metrics are precision and recall. The formula for calculating precision and recall are shown below:

$$\text{Precision} = TP / (TP+FP) \qquad \text{Recall} = TP / (TP+FN)$$

where TP represents the True Positive, which is the number of actually canceled trips that are predicted as canceled, FN represents the False Negative, which is the number of trips predicted to be uncanceled but canceled, and FP represents the False Positive, which is the number of trips predicted to be canceled but actually uncanceled.

Intuitively, the precision is the actually canceled trips among all the trips predicted to be canceled, while the recall is the correctly predicted canceled trips among all canceled trips. We prefer precision as our evaluation metric because making correct predictions and making fewer errors is important in this case. To retrieve an error is costly for RTS since they need to run an extra bus to pick the customer they missed. Meanwhile, recall is also valuable because the RTS wants to capture as many canceled cases as possible.

4.2 Baseline Model: Random Forest Classifier

Our baseline model used for prediction is a random forest classifier. Random forest is a classifier that contains a number of decision trees on various subsets of the given dataset and takes the average to improve accuracy[5]. It consists of a set of decision trees based on randomly selected subsets of the training set and then collects the votes from different decision trees to decide the final prediction[6]. Training a random forest takes a relatively low time and gives high accuracy for a dataset with high dimensions, so we decided to use the random forest classifier as our first attempt.

To optimize the performance, we applied grid search to tune the model. The model with 90 estimators achieved an accuracy of 81.5% and performed best among all the tested models. The precision is 32.2% and the recall is 53.9%. Clearly, the extremely low precision is not desired by our sponsor.

4.3 Class imbalance

One possible cause for the low precision of the model is the class imbalance in our dataset. As shown in Fig. 10, only 19.7% of the trips in the dataset are canceled, which means that even simply guessing all trips will not be canceled gives an accuracy of 80.3%. Therefore, we had to explore some advanced techniques to deal with the class imbalance issue.

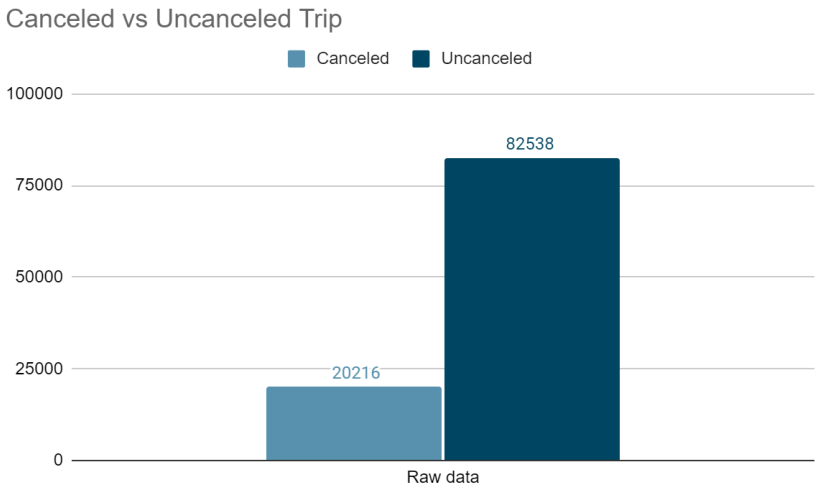


Fig. 10 Distribution of Raw Data

4.4 Model Enhancement: Random Forest Classifier with SMOTE

To solve the class imbalance problem, we first applied the data level approach– Synthetic Minority Oversampling Technique (SMOTE). Synthetic Minority Oversampling Technique is an oversampling technique that synthesizes new examples of the minority class, which is the canceled trips in this project, in the training set. It randomly selects examples that are close in the feature space, drawing a line between the examples and generating a new sample at a point along that line. When the new synthetic examples from the minority class were created, the data in the training set was no longer imbalanced. As shown in Fig. 11, the canceled samples in the training set increased from 16,210 to 66,002.

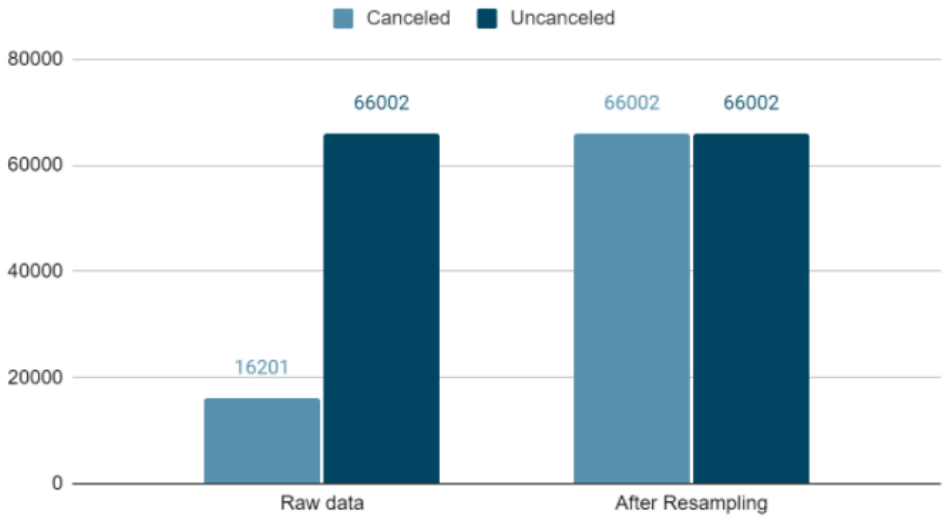


Fig. 11 Distribution of training set before and after oversampling by SMOTE

Our next step was to incorporate this technique with our baseline model. By training a new random forest model with the oversampled training set, we achieved a better outcome testing the model in the test set: the accuracy increased from 81.5% to 84.8%, the precision increased from 32.2% to 62.1%, and the recall increased from 53.9% to 57.7%.

4.5 Model Enhancement: Weighted Random Forest Classifier

Instead of using resampling techniques, we also came up with another strategy for solving the class imbalance issue. Specifically, we applied the Weighted Random Forest Classifier. Compared to a simple Random Forest Classifier, a Weighted Random Forest Classifier assigned a higher weight to the minority class. Whenever it misclassified the minority class, the loss function of the classifier would be multiplied by a higher class weight. The intuition behind this was that the Weighted Random Forest Classifier penalizes misclassification of the minority class more than the Random Forest Classifier.

We built the Weighted Random Forest Classifier by setting the class_weight argument of the Random Forest Classifier as ‘balanced’. The argument value of ‘balanced‘ automatically used the inverse weighting from the training dataset, focusing on the canceled trips.

The result suggested that the Weighted Random Forest Classifier improved the precision significantly, which achieved 75.1%. However, as a trade-off, the recall slightly decreased to 45.1%. The overall accuracy of the model increased to 86.4%.

4.6 Model Enhancement: XGBoost Classifier

Besides the data resampling techniques and the variation of the Random Forest Classifier, we explored Algorithmic Ensemble Techniques as our final step–the Extreme Gradient Boosting(XGBoost) Classifier. Boosting is a type of ensemble learning that uses the previous model's result as an input to the next one. The XGBoost Classifier randomly built an initial model that predicted the cancellation of the trips and calculated the error for each observation in the dataset. It further made predictions on the error based on the Decision Tree Classifier, included the prediction into the ensemble of models, and repeated this process.

To optimize the performance of the XGBoost Classifier, we also performed parameter tuning by grid search. We tuned the parameters, including the learning rate(learning_rate), the number of trees in the ensemble(n_estimators), and the maximum depth of a tree(max_depth). The optimized performance is found when learning_rate=0.11, n_estimators=1000, and max_depth=10. Moreover, to overcome the class imbalance, we adjusted the balance of positive and negative weights(scale_pos_weight) to 4, which is the fraction of canceled instances divided by the uncanceled instances.

After choosing the best hyper-parameters, we evaluated the test performance. The XGBoost Classifier achieved a balance between precision and recall, where the precision increased to 65.2% and the recall reached 62.3%. The overall accuracy of the model also increased to 86.1%.

4.7 Performance and Result

After exploring different models, we compared the performance of each model. Here we have used some abbreviations for some models: ‘RF’ denotes the Random Forest Classifier. ‘SMOTE’ denotes the Synthetic Minority Oversampling Technique, and ‘XGBoost’ denotes Extreme Gradient Boosting.

Table. 4 Test Performance for Different Models

Models	Recall	Precision
RF	53.9%	32.2%
RF + SMOTE	57.7%	62.1%
Weighted RF	45.1%	75.1%
XGBoost	62.3%	65.2%

As we can tell from the table, the Weighted Random Forest Classifier provided the highest precision while the XGBoost Classifier had the most balanced recall and precision.

5. Conclusion & Future work

In conclusion, after simple preprocessing of the RTS dataset and weather dataset, we first conducted exploratory data analysis and found that there are some relationships between cancellation and different variables such as trip type, trip purpose, requested trip time, snowfall, and so on. Then we built four classification models based on the information provided in the reservation to predict the possible cancellation. We found that the Weighted Random Forest Classifier provided the highest precision while the XGBoost Classifier had the most balanced recall and precision.

Incorporating a business perspective, we concluded that the choice of model should be case by case. First, our sponsor, RTS, always has an extra bus on standby to cover any missed cases. Maintaining the extra bus is costly even if there is no missed case. If we appropriately take advantage of the extra bus by having a small number of reserved trips covered by it, the RTS can make a more productive and less costly schedule using the predictive model. Therefore, precision should be valued more during busy hours, since a model with higher precision makes fewer errors and during busy hours, making mistakes costs more due to having many requests. We decided to use the Weighted Random Forest Classifier to predict the cancellation during busy hours, with a precision of 75.1% and a recall of 45.1%. In a non-busy case, a balanced recall and precision model is the optimized solution, where we chose the XGBoost Classifier with a precision of 65.2% and a recall of 62.3%.

There are a few challenges we encountered during the modeling procedure:

1. There is a significant class imbalance issue in the dataset, that the initial dataset contains 80% "uncanceled" trips, whereas the "canceled" trip only accounts for roughly 20%. This causes the first-round model prediction to be inaccurate due to low precision and the model accuracy less reliable.
2. There is a tradeoff between "making precise predictions" and "covering more cancellation data". When we were building the model, balancing precision and recall was another obstacle we dealt with.
3. The weather data we collected is limited to the time range from May to December. A dataset with a larger time span is required to make more generalized conclusions.

Some possible future improvements and further steps of this project could be taken, as we are considering a training model with a wider time span of internal and external data. Also, for forecasting the regional transportation service, some geographical analysis techniques could help explain the routes and locations.

References

[1] UKEssays. (November 2018). Effectiveness of Para-Transit Transport Services. Retrieved from https://www.ukessays.com/essays/geography/effectiveness-paratransit-transport-9681.php?vref=1
[2] Ali Tizghadam, Hamzeh Khazaei, Mohammad H. Y. Moghaddam, Yasser Hassan, "Machine Learning in Transportation", Journal of Advanced Transportation, vol. 2019, Article ID 4359785, 3 pages, 2019. https://doi.org/10.1155/2019/4359785
[3] Zackowitz IB, Vredenburgh AG. When Communication Failure Contributes to an Injury: A Case Study of Para-Transportation for Wheelchair Users. Proceedings of the Human Factors and Ergonomics Society Annual Meeting. 2007;51(9):559-563. doi:10.1177/154193120705100902
[4] G. Meena, D. Sharma, and M. Mahrishi, "Traffic Prediction for Intelligent Transportation System using Machine Learning," 2020 3rd International Conference on Emerging Technologies in Computer

Engineering: Machine Learning and Internet of Things (ICETCE), 2020, pp. 145-148, doi: 10.1109/ICETCE48199.2020.9091758.
[5] https://www.javatpoint.com/machine-learning-random-forest-algorithm
[6] https://www.geeksforgeeks.org/random-forest-classifier-using-scikit-learn/