Investigator(s) Information

Sheth	Pal	40196640
Safavi Sohi	Seyed mohammadhossein	40199128
Sheth	Rajvi Maheshbhai	40211702
Vaghasiya	Vikas Vipulbhai	40219501
Prajapati	Jay	40227110

Project's General Information

Title:	**Prediction of the bike availability and uneven distribution of BSS using Machine Learning.**		
Infrastructure Sector	Transportation	**Sub-sector:**	Efficiency.
Municipality:	City of Chicago	**Region:**	USA

Abstract- *The major issues with the BSS system are uneven bike distribution and user dissatisfaction. The growing number of bike-sharing trips leads to uneven distribution and user dissatisfaction. Because of user dissatisfaction, the funding opportunity in BSS has been significantly reduced, resulting in a financial crisis in the organization. This study aims to forecast the number of bikes at the station, which will aid in understanding the uneven distribution, resource allocation, and user engagement. External factors such as temperature, humidity, and weather events such as rain, sun, and snow heavily influence user engagement and usage of BSS. There are many bike tracking systems available, but there are no applications for owners to track real-time and future demand, which shows the departure and arrival of bikes at each station. In this paper, we used machine learning algorithms to forecast the number of bike departures and arrivals at each station while accounting for weather and hourly time distribution. The study's objectives include (1) investigating the attributes that are correlated to departure and arrival, and (2) developing a prediction model for predicting the number of departures and arrivals at a specific station based on the correlated attributes. For this study, three classification models were employed: logistic regression, decision tree, and k-nearest neighbour (kNN). Based on the results, logistic regression performed well in binary classification, while kNN outperformed the decision tree.*

1. Introduction & Background

In many cities worldwide, the bike-sharing system (BSS) is becoming a more common mode of transportation in urban areas, with China leading the way. The project of choice for urban mobility is ridesharing. According to a statistical study, the USA has launched several bike-sharing systems in its city. The total number of transportation options in the USA grew in 2018. According to NACTO, around 2.4 million journeys were shared in 2012, and 84 million were in 2018, which increases exponentially. BSS has some unique characteristics like environmentally friendly, less traffic congestion, and easy access which makes it different from other available systems like Public Transport, Car Pooling, and Car Sharing [1]. Recently, station-based bike-sharing systems, which allow users to pick up and drop off bikes at stations, have gradually been evolving. According to the Bike Sharing World Map, as of March 2019, there were approximately 1950 operational bike-sharing programs and approximately 14,860,200 bikes in service worldwide [2]. With this expansion, BSS has grown considerably and has become a popular mode of travel for

individuals. Even as demand has increased, BSS's efficiency has not scaled well. One of the big reasons is the insufficient number of bikes at stations problem in BSS, which can be framed as a rebalancing problem [1]. The motivation to focus on this problem is to improve efficiency and user engagement.

2. Problem Statement and Objective

Pronto Company is a bike-sharing service provider that operates in the city of Seattle, USA. Users can borrow and return bikes at one of their many stations located across the city. For customer satisfaction and user engagement, the company must ensure that each station has an adequate quantity of bikes available for users because the demand for bikes varies throughout the day. The uneven distribution of bikes and the availability of bikes at dock stations in peak hours play a major role in user satisfaction and revenue generation. The application of machine learning will help in the identification of the uneven distribution and weather impact on rides to make the system more efficient and available for the user.

The BSS company wants to know how many bikes will be added or reduced in the next hour in a specified station. Knowing that we can manage our equipment and bike-carrying trucks to refill the stations with no bikes.

For this to achieve we have three main objectives.

1. Exploring the correlation between the different attributes such as weather, time, and date on the number of trips in a particular time.
2. Predicting a number of departures at a particular time at a particular station based on correlated attributes, using different machine-learning algorithms.
3. Predicting a number of arrivals at a particular time at a particular station based on correlated attributes, using different machine-learning algorithms.

3. Previous works

Demand forecasting, inventory decision-making, and rebalancing are the primary operational-level decision problems of bike-sharing systems. Each of these issues is typically addressed in the order listed, with inventory targets serving as constraints for the rebalancing (routing) problem and demand forecasts serving as inputs to determine these inventory targets. "Bike-sharing demand is known to be heavily dependent on temporal information (intra- and inter-day) but also on the weather" [3].

[4], The purpose of this paper was to identify the most significant macro-environmental factors influencing the long-term development of bike-sharing systems (BSSs) in Polish cities. Expert methods, STEEP environmental assessment, factor evaluation with a weighted score, and fundamental analysis method with MICMAC computer programmed were used in the study to analyze and categorize the key factors influencing the development of a BSS into five distinct categories. According to the literature, different external environmental factors affect the creation and long-term viability of public bike-sharing systems. These factors include social, technological, economic, environmental, and political factors, all of which interact to determine whether BSSs succeed or fail.

[5], the study offers an innovative method for predicting bike availability in BSS using machine-learning algorithms. Random Forest outperforms Least-Squares Boosting in univariate modelling

at the station level, while Partial Least-Squares Regression (PLSR) is helpful for multivariate modeling at the network level. Bike counts are significantly predicted by demographic information and environmental factors. The study reveals that most trips in the San Francisco BSS were short-distance trips, which may affect the application of overtime fees for trips lasting more than 30 minutes. Furthermore, station Neighbours and a prediction horizon time of 15 minutes were discovered to be significant predictors, with longer prediction horizons resulting in higher error rates.

In 2022, [6] proposed a method in which they used the data of Citi bikes from the city of New York. Their primary objective was to make a model to evaluate the station-level pickup and return demand in BSS and use that decision for starting-level inventory management. They used 30 active station data for 1 year i.e., from Jan 2018 till Dec 2018 for the model training and evaluation. Individual records of users renting and returning bikes comprise the data, which we aggregate into three distinct temporal aggregation levels: 15-, 30-, and 60-min intervals. And additional weather data has been considered for better prediction. They analyze the performance of the proposed model i.e., VP-RNN on the task of pickup and return bike-sharing demand prediction, and then compared the performance of the proposed VP-RNN with other learning-based approaches including historical average, Moving average, Linear regression, Gaussian RNN, Poisson RNN. The study's findings indicate that forecasts and decision models should be thoroughly assessed and unified to effectively control shared mobility systems.

[7], the paper proposes a technique for improving the management and operation of shared bicycle systems (BSS) in urban environments by combining IoT and machine learning. The simulation results show that the combined methods are more effective than the individual methods resulting in appropriate recommendations for system stakeholders. The potential of connected technologies in BSS for improving urban mobility and promoting sustainability is highlighted. This paper adds to the growing body of literature on the application of IoT and machine learning in the management and optimization of urban transport systems.

[8] investigated different variables influencing BSSs in over 50 cities. He assessed the efficiency and service quality of BSSs using an analysis of benchmarks based on Key Performance Indicators (KPIs) and customer satisfaction. His research offers an intriguing statistical evaluation of BSS data as well as some insights into the related business models.

BSS was used to create a personal journey advisor for navigating in Dublin. [9] they used a spatiotemporal forecasting system based on Auto-Regressive Integrated Moving Average (ARIMA) to forecast the number of bikes in a station in the near future (5 and 60 minutes ahead), which additionally takes variations in the seasons and spatial correlations into account. Their application suggests the best pair of stations to take and return a bicycle based on the origin and destination.

[10] proposed a dynamic cluster-based framework for BSS over-demand prediction. They built a weighted correlation network based on contextual factors like time, weather, social, and traffic events to group stations into clusters with similar usage patterns. They proposed a Monte Carlo simulation to predict each cluster's over-demand probability. They demonstrated that their framework could accurately predict over-demand clusters by applying the proposed model to real-world data from New York City and Washington, D.C.

[11] utilized a linear regression model to create a model of cyclic temporal patterns, which they then used to forecast. They made use of factors such as weather, user count, and holiday indicators. Their research reveals the spatial and temporal patterns of activity in the city, as well as hourly or daily forecasts of available bikes. They examined data from Velo'v, Lyon, France's shared bicycle programmed.

[12] modelled bicycle traffic in Vancouver, Canada, using continuous and year-round hourly bicycle counts and weather data. A seasonal autoregressive integrated moving average analysis was used in that study to account for the complex serial correlation patterns in the error terms, and the model was tested against actual bicycle traffic counts. The findings revealed that the weather had a significant and significant impact on bike usage. The authors discovered that weather data (specifically temperature, rain, humidity, and clearness) were generally significant; temperature and rain had a significant effect.

4. Methodology

4.1 Data Understanding

The data was obtained from the Kaggle open data source provided by the Seattle government. In the project, a data warehouse has been created from three datasets: Trip data, Weather data, and Station data, which consisted of 500 bikes and 54 stations located in Seattle by the Pronto Cycle Share System. The project data from March 2014 to August 2016 has been considered, in which trips start time-end time, and weather forecasts of each day have taken place.

Each of the datasets listed above has several attributes, and the most frequent and significant ones are chosen as the attributes of our case study based on the attributes that are most frequently mentioned in the literature. The table lists each attribute's name, type, and other details.

Table 1: Attribute description

Attribute	Description	Data type
Date	The date of the bicycle trip.	Date and time
Temperature_F	The temperature that day, in Fahrenheit.	Numerical
Humidity	The relative humidity on the day of the trip.	Numerical
Precipitation_In	The amount of precipitation in inches on the day of the trip.	Numerical
Events	Any weather-related incidents (e.g., rain, snow, etc.).	Nominal
station_id	The ID of the bike share station where the trip started or ended.	Nominal
name	The name of the bike share station.	Nominal
lat	The latitude of the bike share station.	Numerical
long	The longitude of the bike share station.	Numerical
install_date	The date when the bike share station was installed.	Date and time
install_dockcount	The number of bike docks at the time of installation	Numerical
current_dockcount	The current number of bike docks	Numerical
trip_id	The ID of the bike trip.	Nominal
starttime	The date and time when the trip started.	Date and time
stoptime	The date and time when the trip Ended	Date and time
bikeid	The ID of the bike that was used for the trip.	Nominal
tripduration	The duration of the trip in seconds.	Numerical
from_station_id	The ID of the bike share station where the trip started.	Nominal
to_station_id	The ID of the bike share station where the trip ended.	Nominal
usertype	"Short-Term Pass Holders" - 3 to 24 hours duration pass "Members" refer to monthly or annual pass	Nominal
gender	The gender of the user who made the trip.	Nominal
birthyear	The birth year of the user who made the trip	Numerical

4.2 Data Preparation

The information was provided in two Excel files, each containing approximately 250,000 records. Because data contains missing or inconsistent values, data cleansing is required to obtain a better model; thus, we begin our data preparation by filtering missing values and removing inconsistent data, as illustrated in the sections Missing Data and Inconsistent Data. change the format of this paragraph and increase the content.

4.2.1 Data warehouse

The data warehouse is the process of combining data from various sources to create a single set of data. In our case, we used the join operator to create a data warehouse of weather and trip data. Because we want to find departure trips from the station, we chose the left key attribute as the start date for the departure case, whereas the end date is used for the arrival case.

4.2.2 Converting the raw trip data into hour format.

The main source data showed all the trips independently, so we created a pivot table in Excel and used it to convert all the trips into specific hour based, which helped us in our model to input as the main CSV file to predict the departure and arrival for each hour at each station.

trip_id	starttime	stoptime	bikeid	tripduratio	from_station_name	to_station_name	from_station_	to_station_i	usertype	gender	birthye
431	10/13/2014 10:31	10/13/2014 10:48	SEA00298	985.935	2nd Ave & Spring St	Occidental Park / Occidental Ave S & S Washington St	CBD-06	PS-04	Member	Male	1960
432	10/13/2014 10:32	10/13/2014 10:48	SEA00195	926.375	2nd Ave & Spring St	Occidental Park / Occidental Ave S & S Washington St	CBD-06	PS-04	Member	Male	1970
433	10/13/2014 10:33	10/13/2014 10:48	SEA00486	883.831	2nd Ave & Spring St	Occidental Park / Occidental Ave S & S Washington St	CBD-06	PS-04	Member	Female	1988
434	10/13/2014 10:34	10/13/2014 10:48	SEA00333	865.937	2nd Ave & Spring St	Occidental Park / Occidental Ave S & S Washington St	CBD-06	PS-04	Member	Female	1977
435	10/13/2014 10:34	10/13/2014 10:49	SEA00202	923.923	2nd Ave & Spring St	Occidental Park / Occidental Ave S & S Washington St	CBD-06	PS-04	Member	Male	1971
436	10/13/2014 10:34	10/13/2014 10:47	SEA00337	808.805	2nd Ave & Spring St	Occidental Park / Occidental Ave S & S Washington St	CBD-06	PS-04	Member	Male	1974
437	10/13/2014 11:35	10/13/2014 11:45	SEA00202	596.715	Occidental Park / Occider	King Street Station Plaza / 2nd Ave Extension S & S Jacks	PS-04	PS-05	Member	Male	1978
438	10/13/2014 11:35	10/13/2014 11:45	SEA00311	592.131	Occidental Park / Occider	King Street Station Plaza / 2nd Ave Extension S & S Jacks	PS-04	PS-05	Member	Male	1983
439	10/13/2014 11:35	10/13/2014 11:45	SEA00486	586.347	Occidental Park / Occider	King Street Station Plaza / 2nd Ave Extension S & S Jacks	PS-04	PS-05	Member	Female	1974
440	10/13/2014 11:35	10/13/2014 11:45	SEA00434	587.634	Occidental Park / Occider	King Street Station Plaza / 2nd Ave Extension S & S Jacks	PS-04	PS-05	Member	Male	1958

Figure 1: Raw Data Set

startdate	Time_24	Day_Text	Date	Month	Year	BT_01	....	BT-03	BT-04	Grand T
10/13/2014	10	Monday	13	10	2014		....			6
10/13/2014	11	Monday	13	10	2014		....			54
10/13/2014	24	Monday	13	10	2014		....	2	3	42
10/13/2014	13	Monday	13	10	2014		....		2	39
10/13/2014	14	Monday	13	10	2014		....		1	56
10/13/2014	15	Monday	13	10	2014	3	....			33
10/13/2014	16	Monday	13	10	2014		....	2	2	48
10/13/2014	17	Monday	13	10	2014	3	....		1	48
10/13/2014	18	Monday	13	10	2014	4	....	1	1	42
10/13/2014	19	Monday	13	10	2014		....			12

Figure 2: Hourly Trip Data

4.2.3 Data cleansing (Remove irrelevant data and remove missing data)

The specific hour in which there was no single trip was removed after converting from raw data to the required form of the data. As a result, the previous trip data contains approximately 230,000 data points, which has been reduced to approximately 14,500 data points in hourly format. In addition, the station IDs 8D OPS 02, Pronto Shop 2 and Pronto Shop were incorrect in the data and were removed from the data points. Also, the station without any trips for the particular hour is removed.

4.2.4 Outliers

We used an area diagram in data visualization to detect attribute outliers and made assumptions based on the literature paper. The y-axis represents the total number of trips, and the x-axis represents the number of hours. It was assumed here that trips between 7:00 PM and 7:00 AM are

not used in the model to predict departure and arrival at specific stations. We considered all trips between these hours to be outliers because the frequency of trips between these hours is very low. Because there were more than 15 trips from one station in a very rare instance, we decided to classify the extra trips as an anomaly to reduce the unevenness in the data.

4.2.5 Missing Data

There are some data points in the attribute temperature that have missing values, so we took the average of the temperatures at those locations. The weather condition at the location of missing events is sunny, according to the attribute named events. At a specific time when another station had a certain number of trips and this station had none, we replaced the missing place in trips with zero.

4.2.6 Data Normalization

Numeric data attributes should be normalized for distance-based techniques like KNN, which we used in this study. This process is used to avoid scale problems with some attributes that have different ranges.

4.3 Data Visualization

Data visualization is a crucial step in data analysis to broadly understand the data points. We have analyzed the trip data hourly (Fig. 3) and monthly (Fig. 4). The graph between trips and hours represents the most significant hours of the day and helps in identifying those hours which are not relevant for further studies, i.e., 7 AM to 7 PM hours are considered for the analysis of the data.

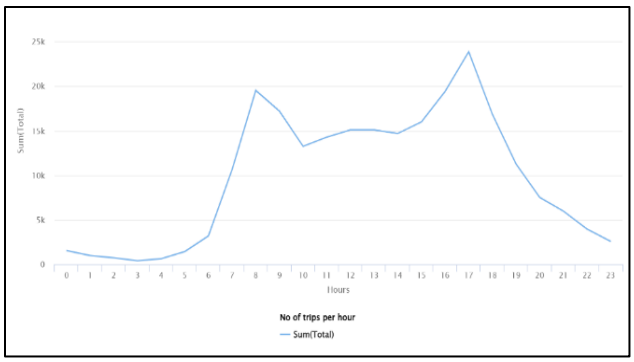

Figure 3: Number of trips per hour

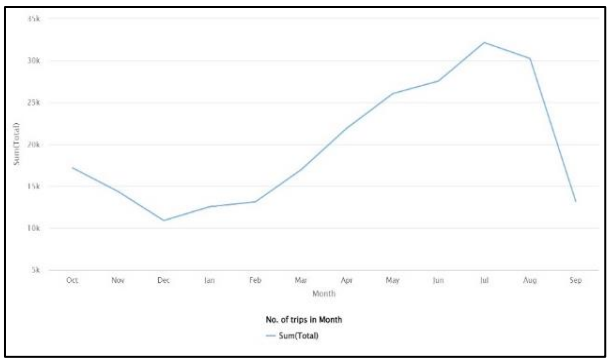

Figure 4: Number of trips per month

The monthly distribution of the trips indicates the maximum number of trips and the significance of the temperature on the trips since the weather is directly related to the month. As per the graph, there is a declining trend during the month of Oct to March i.e., during the winter and on the other side trips are increasing during the summer. Additionally, we have also analyzed the impact of the temperature on the total number of trips, from the graph below, we observed that with an increase in the temperature, there is an increase in the number of trips i.e., direct relation between temperature and the number of trips.

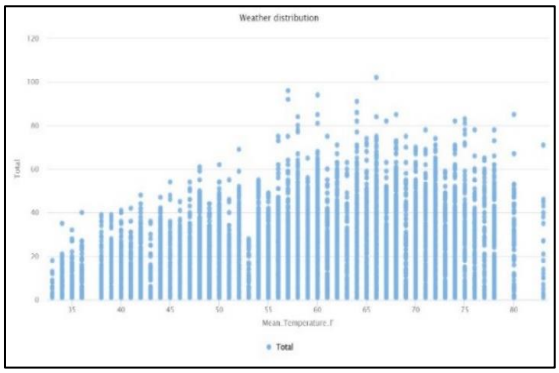

Figure 5: Cumulative trips at temperature variation

4.4 Pearson Correlation Analysis

Correlation analysis was used in this study because the data majority was in numeric format, which is normalized, and decision trees don't make assumptions about the relationships between features. Based on an impurity measure, it merely splits single attributes that aid classification. If attributes A and B have a strong correlation, splitting on B after splitting on A will yield little to no information. It would therefore typically be disregarded in favour of C, the label. Knowing this helps us to better comprehend why certain attributes were not present in our decision tree. In this study, no significant correlation was found, but some relationships between a few attributes, including mean temperature, humidity, and sunny event, were identified and used together with the labelled attribute total number of trips for further analysis. We have formed a Pearson correlation matrix for both arrival and departure identically. For the Pearson correlation, we have used dummy coding for the Events since all the Events are polynomial attributes.

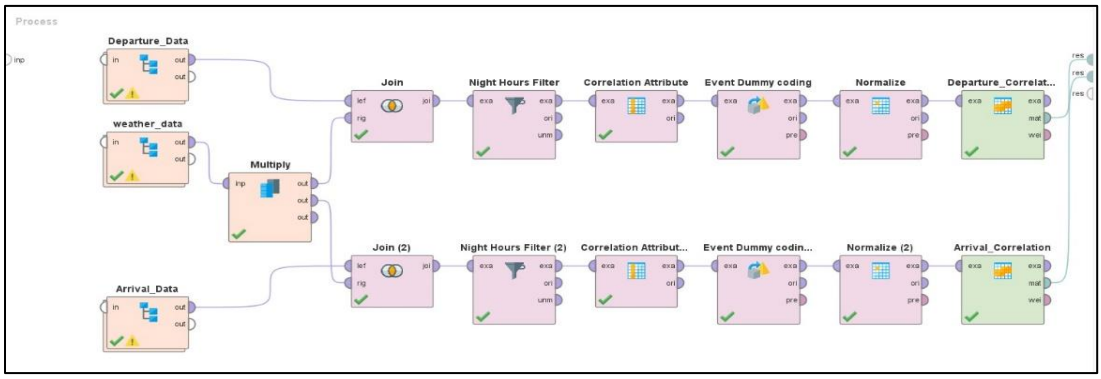

Figure 6: Pearson correlation process

Attributes	3_Event = Rain	3_Event = Sunny	3_Event = Other	Time_24	Date	Month	Grand Total	Mean_Temp...	Mean_Humidity
3_Event = Rain	1	-0.888	-0.211	0.002	0.030	0.014	-0.267	-0.284	0.508
3_Event = Sunny	-0.888	1	-0.266	-0.002	-0.015	0.012	0.300	0.381	-0.589
3_Event = Other	-0.211	-0.266	1	-0.001	-0.031	-0.053	-0.076	-0.213	0.207
Time_24	0.002	-0.002	-0.001	1	0.002	-0.001	0.167	-0.008	0.006
Date	0.030	-0.015	-0.031	0.002	1	0.024	-0.025	0.008	0.022
Month	0.014	0.012	-0.053	-0.001	0.024	1	0.029	0.108	-0.143
Grand Total	-0.267	0.300	-0.076	0.167	-0.025	0.029	1	0.439	-0.403
Mean_Temperature_F	-0.284	0.381	-0.213	-0.008	0.008	0.108	0.439	1	-0.588
Mean_Humidity	0.508	-0.589	0.207	0.005	0.022	-0.143	-0.403	-0.588	1

Figure 7 Pearson correlation matrix for arrival

Attributes	3_Event = Rain	3_Event = Sunny	3_Event = Other	Time_24	Date	Month	Grand Total	Mean_Temp...	Mean_Humid...
3_Event ...	1	-0.887	-0.210	0.002	0.031	0.015	-0.274	-0.285	0.508
3_Event ...	-0.887	1	-0.266	-0.001	-0.016	0.010	0.306	0.381	-0.598
3_Event ...	-0.210	-0.266	1	-0.001	-0.032	-0.052	-0.078	-0.212	0.206
Time_24	0.002	-0.001	-0.001	1	0.000	-0.003	0.090	-0.009	0.004
Date	0.031	-0.015	-0.032	0.000	1	0.024	-0.025	0.008	0.021
Month	0.015	0.010	-0.052	-0.003	0.024	1	0.030	0.107	-0.142
Grand To...	-0.274	0.306	-0.076	0.090	-0.026	0.030	1	0.457	-0.414
Mean_Te...	-0.285	0.381	-0.212	-0.009	0.008	0.107	0.457	1	-0.595
Mean_Hu...	0.508	-0.598	0.206	0.004	0.021	-0.142	-0.414	-0.595	1

Figure 8: Pearson correlation matrix for departure

The strength and direction of the correlation between two variables are shown by colour-coded squares on a heat map, which is a graphical representation of correlation matrices. It is frequently used to visually discover relationships and patterns among numerous variables. On the basis of the literature review and research, we used it here initially just to look at the correlation between the majority of the selected attributes.

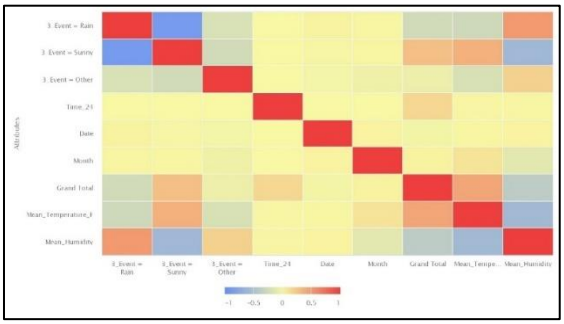

Figure 9: Heat Map

4.5 Modellings

4.5.1 Range Selection

Given that the data we have is unevenly distributed, logistic regression is first used to eliminate the value of 0, which denotes the absence of any trips. After that, the range of data was created for the remaining data. First, if there have been one or two trips, and if there have been more than three, there are two groups. Additionally, the number of trips was kept constant when the second range was created, with the exception of those groups containing more than ten trips. Trial and error were used to determine the range distribution for trip occurrence, with model F1 score and accuracy being key considerations.

4.5.2 Grid Optimization Operator

To find the optimum value of different parameters such as the number of folds, nearest neighbours and a number of neighbours in SMOTE sampling, the optimize parameters (Grid) operator is used, in which set parameters configure operators the number of different parameters is selected and value of minimum, maximum and step size for the grid is assign having linear scaling in such a manner to receive the best number of value of the selected parameter. For example, in Fig. 10 we have used the grid optimization tool for calculating the optimum number of folds required for cross-validation, similarly, the tool has been utilized for the identification of the value of k in kNN and the number of neighbours in SMOTE for oversampling of the unbalanced data set.

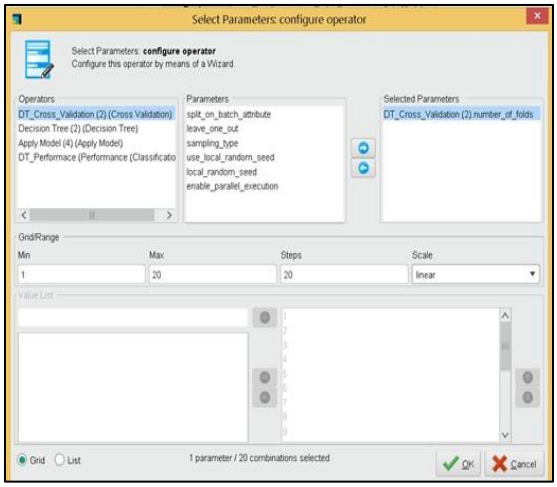

Figure 10: Optimization for number of folds in Cross Validation

4.5.3 Logistic Regression

Logistic regression is a statistical technique used for binary classification problems, such as predicting whether a certain event has occurred or not. In this case, the logistic regression model is used to predict whether a trip has been taken or not, with the outcome variable taking on the value of 0 if the trip has not been taken and 1 if the trip has occurred. To run the process, we have converted all the trips into 1 i.e., if a trip has occurred at one station, then it will be counted as 1 or else 0.

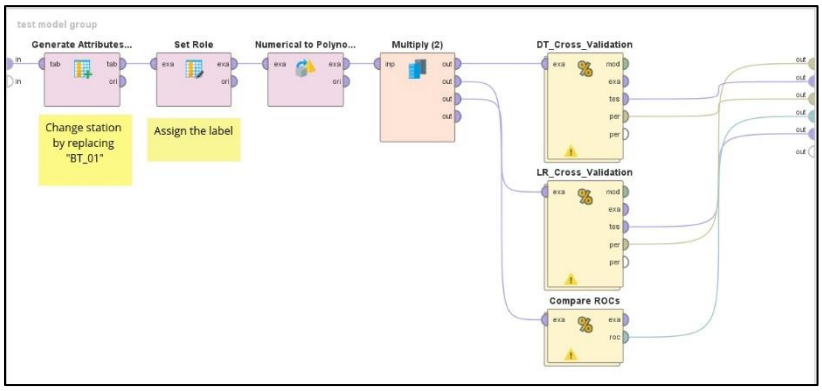

Figure 11: Process of Logistic Regression

4.5.4 Decision Tree

The problem belongs to the classification set of prediction models and the Decision Tree classifier was chosen to predict the number of arrivals and departures. The decision tree method is primarily used because it can explicitly and visually represent the prediction process and various generated rules. To find the best attribute in each node by reducing entropy, Information Gain was set to 0.01 as the criterion. It was decided to pre-prune with a minimal gain of 0.01 and prune with a confidence level of 0.01. Since DT is extremely sensitive to overfitting, it is crucial to choose the right depth, setting, and pre-pruning. To avoid overfitting, we should be cautious after experimenting with various depths and using a grid optimization tool to find the appropriate depth. The maximum depth of 15 was chosen as a result. To implement the Decision Tree model, all missing values and attributes were filtered out, as shown in Figure 5. The departure station was designated as the target variable, and the range was divided into two groups: Group 1: 1 or 2 trips; Group 2: 3 or more trips.

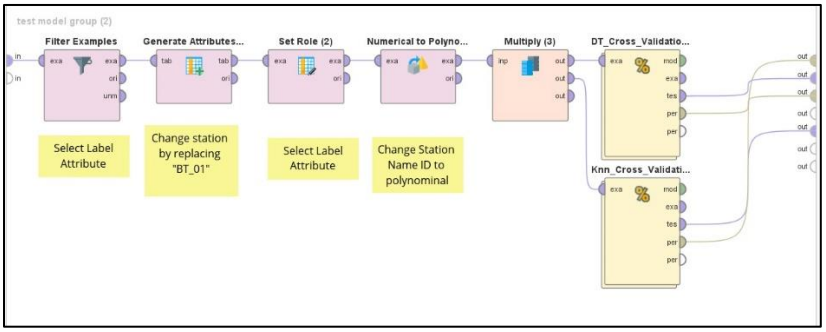

Figure 12: Process of Decision Tree

4.5.5 k Nearest Neighbor (kNN)

The k-nearest neighbour (kNN) model has been developed for the range 3 with the 7 folds. This 7-fold has been finalized using the grid optimizer tool. Now, Attribute - BT_01 Has been labelled as a targeted attribute in the model development process. As kNN cannot perform in the non-nominal data, the dataset has been converted to the nominal data. After getting the nominal data, the synthetic data was created using the SMOTE up sampling tool in the RapidMiner, but all the ten groups were supposed to have an equal amount of the synthetic data to achieve that value; SMOTE has been looped ten times with the help of generate attribute operator to catch the updated synthetic data in each iteration. After having the equal value for each group, the addition generated attribute was removed so that the dummy effect could not produce.

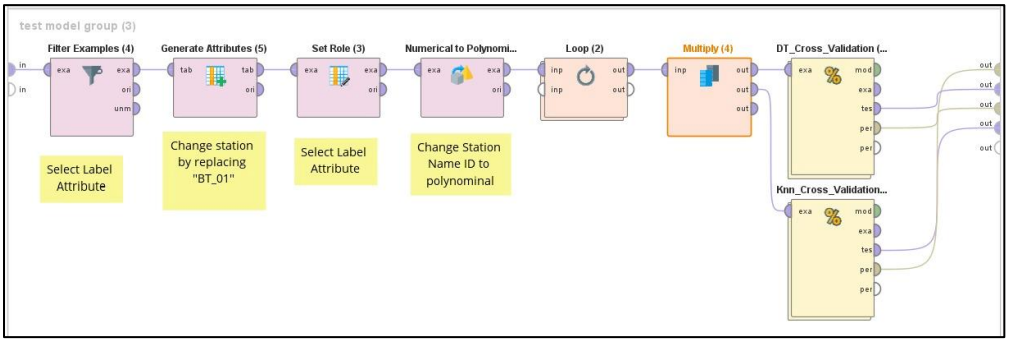

Figure 13: Process of kNN

5. Evaluation

To evaluate the result of the binary classification model we used the "Compare ROC" operator to choose the final model. Since the ROC curve cannot be used for the final decision of the model because of its limitation, we have compared the final output results of all the models to decide and use the data for the next model. Our aim of this binary classification was to identify whether the trip has occurred or not so the model which is able to predict a more accurate zero is considered as the best model for binary classification.

In our case, the ROC of the decision tree was found to be more a predictive (Fig. 14 & 15) but its accuracy in predicting the zero (Recall) bikes is much less than that of the logistic curve, hence in the finalization of the model we have considered the output of the logistic regression as a benchmark since its recall value in predicting zero is more than that of the decision tree see Table 2 of the confusion matrix. For the given case we have not included the confusion matrix of the kNN because of its underperformance.

We have also compared the AUC for all four cases which are given in Table. 2 which implies that the Decision tree is good for predicting both cases, but we are not interested in predicting the 1, i.e., the case where a trip has occurred. Hence even though the F1 score is more for the decision tree we have used logistic regression as a final model.

AUC is calculated as.

$$\text{AUC} = \frac{\text{Precision} + \text{Recall}}{2}$$

Table 2: AUC comparison

AUC	Model	Departure	Arrival
Binary Classification	Logistic Regression	60.97%	62.70%
	Decision Tree	65.41%	64.80%

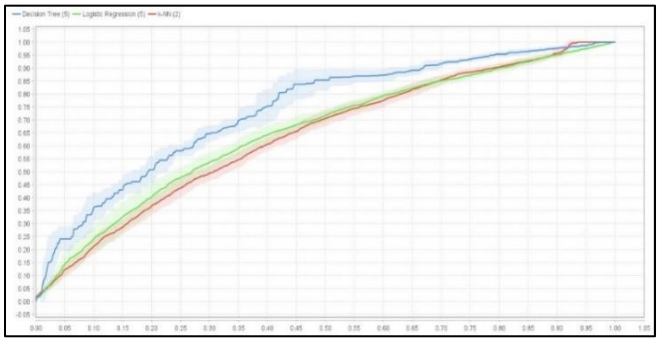

Figure 14: ROC of departure at BT_01

Table 3 Confusion matrix of departure (Logistic regression) & Confusion matrix of departure (Decision tree)

accuracy: 67.02% +/- 1.94%			
	true 0	true 1	class precision
pred. 0	3898	1550	71.55%
pred. 1	1321	1937	59.45%
class recall	74.69%	55.55%	

accuracy: 64.12% +/- 1.74%			
	true 0	true 1	class precision
pred. 0	4402	2307	65.61%
pred. 1	817	1180	59.09%
class recall	84.35%	33.84%	

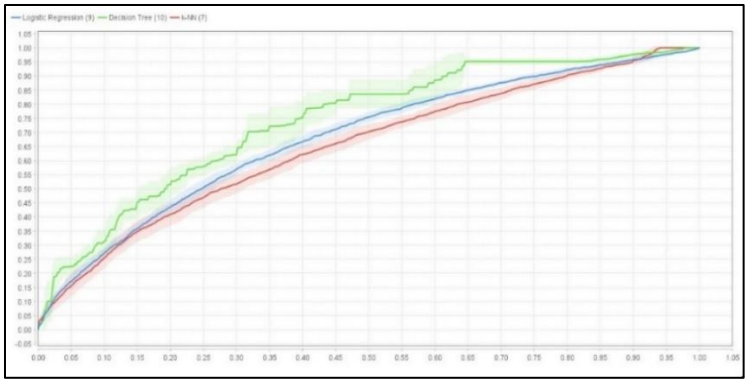

Figure 15: ROC of arrival at BT_01

Table 4 Confusion matrix of arrival (Logistic regression) & Confusion matrix of arrival (Decision tree)

accuracy: 65.06% +/- 1.12%			
	true 0	true 1	class precision
pred. 0	4167	2079	66.71%
pred. 1	970	1510	60.89%
class recall	81.12%	42.07%	

accuracy: 66.42% +/- 1.42%			
	true 0	true 1	class precision
pred. 0	3876	1669	69.90%
pred. 1	1261	1920	60.36%
class recall	75.45%	53.50%	

To check the performance of the 2nd and 3rd model we understood the overall accuracy and calculated the F1 score since the data was unbalanced. F1 score is the harmonic mean of the precision and recall.

F1 score is calculated as.

$$\text{F1 Score} = \frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}$$

Table 5: F1 Score Calculation

Range	Model	Departure	Arrival
Binary Classification	Logistic Regression	65.31%	62.68%
	Decision Tree	60.68%	64.80%
Range 2	kNN	57.70%	58.30%
	Decision Tree	53.62%	57.16%
Range 3 (Oversampling)	kNN	71.33%	73.58%
	Decision Tree	57.85%	56.50%

6. Results and discussion

According to the provided accuracy and F1 score comparison table all three models namely Logistic regression, decision tree and kNN are useful for predicting the number of departures and arrival at the individual station. For binary classification, it can be observed that the accuracy of the decision tree is more than that of the logistic regression but its F1 score and recall value is very less for predicting the zero i.e., to predict whether the trip has occurred from one station. Binary classification removes the data which causes the unbalancing of the range. i.e., zero which is

almost 65% of the entire data set, hence predicting zero is our crucial goal for binary classification. Here logistic regression performed very well in predicting zero and hence considered for the evaluation. Furthermore, the occurrence of trip one or more has been divided into different ranges using trial and error, to determine the range distribution for trip occurrence, with model F1 score and accuracy being key considerations.

For the 2nd range either the occurrence of trips, one or two in the first group and 3 or greater in the second group is distributed. As per Table 4, it can be observed that in the accuracy and F1 score of departure and arrival, the kNN model performed better than the decision tree.

For the 3rd range we have removed the zero and all the data set has been kept the same as it is except for 10 or more trips in one group. The issue of uneven distribution of the data has been reduced by an oversampling method using the SMOTE oversampling. For the oversampled data set, it can be seen that the performance of the kNN is very accurate while the decision tree does not perform well. The overall accuracy and F1 score of the kNN are approximately 15% higher than that of the decision tree.

Table 6: Accuracy and F1 score comparison.

Range	Model	F1 Score		Accuracy	
		Departure	Arrival	Departure	Arrival
Binary Classification	Logistic Regression	65.31%	62.68%	64.12%	65.06%
	Decision Tree	60.68%	64.80%	67.02%	66.42%
Range 2	kNN	57.70%	58.30%	71.29%	73.36%
	Decision Tree	53.62%	57.16%	69.43%	72.55%
Range 3 (Oversampling)	kNN	71.33%	73.58%	71.38%	73.31%
	Decision Tree	57.85%	56.50%	58.72%	57.40%

7. Deployment

Bike-sharing system professionals could integrate the method into their existing system to deploy it. This would entail creating an interface that takes weather data and predicts the number of arrivals and departures at each station. This information could then be used by professionals to plan the allocation of bikes and docking stations, as well as other decision-making processes. To ensure the accuracy of the predictions, the data used to train the model must be updated on a regular basis.

7.1 Resource Allocation

The model can provide insights into which stations have higher demand during certain weather conditions. This information can help bike-sharing system professionals allocate their resources efficiently, such as by increasing the number of bikes and docking stations at those stations during those periods.

7.2 Demand Prediction (User Engagement)

By analyzing previous weather data, the model can predict the demand for bikes at a specific station at a particular time of day. Professionals in the bike-sharing system can use this data to plan the availability of bikes and docking stations at various stations which will indirectly increase the user engagement as this will increase the satisfaction of the user.

7.3 Decision-making

The model can provide data-driven insights for decision-making related to bike-sharing system expansion, such as identifying high-demand areas where new stations could be added.

8. Concluding remarks

In this research, we primarily focused on predicting the departure and arrival of trips at individual bicycle stations in Seattle city. We used the open data source from Kaggle and weather data from Seattle. We show that weather conditions and hours plus the month would significantly affect the number of trips. Firstly, we identified the correlated attributes that significantly influence bike prediction. Since the trip data was highly unbalanced, we had two approaches to resolve the issue. First is to remove all the zero, which shows trips did not happen in that hour from that individual station, so to achieve this we used logistic regression to remove all the zero trips from the data whose accuracy is 60.97% and 62.70% for departure and arrival respectively. After that, the rest of the data has been used in the second stage where we use a range or oversampling method to predict the actual number of departures and arrival from that particular station. We used a decision tree and kNN for the prediction, after calculating and comparing the F1 score and average accuracy we concluded that the kNN performs very well in predicting the number of trips. A comparison of the F1 score and accuracy has been given in Table 6.

In this study we faced several issues some can be mitigated using the advanced method, but some cannot be tackled. kNN is a lazy learner and its accuracy gets highly influenced by the outlier. Moreover, finding the value of k is a big task in big data, to overcome these difficulties we have used different approaches including the “optimize parameter (Grid)” operator, outlier reduction using the normalization technique and oversampling. In addition to that several problems such as the total number of bikes and the latest updated data were not available for training the model. We believe that if we train the model with more data and provide the updated data then it can predict the number of departures and arrival with great precision and accuracy which help the company in identifying the location for resource allocation and helps customers in identifying the availability of bikes at a particular station.

9. References

[1] J. P. V. P. B. Bhargav Shir, "Mobility prediction for uneven distribution of bikes in bike sharing systems," Wiley, Gujarat, 2022.

[2] N. Z. S. M. R. L. Chenyi Fu, "A two-stage robust approach to integrated station location and rebalancing vehicle service design in bike-sharing systems," Elsevier, 2022.

[3] V. E. U. Ezgi Eren, "A review on bike-sharing: The factors affecting bike-sharing demand," *elsevier,* p. 12, 2019.

[4] M. C. El˙zbieta Macioszek, "External Environmental Analysis for Sustainable Bike-Sharing System Development," *energies* , p. 22, 2022.

[5] M. E. ,. H. A. R. ,. M. A. ,. a. L. H. Huthaifa I. Ashqara, "Network and station-level bike-sharing system prediction: a San Francisco bay area case study," *Journal of Intelligent Transportation Systems,* p. 13, 2021.

[6] Y. W. D. P. F. R. S. M. F. C. P. Daniele Gammelli, "Predictive and prescriptive performance of bike-sharing demand forecasts for inventory management," *Elsevier,* p. 18, 2022.

[7] S. C. K. T. El Arbi Abdellaoui Alaoui, "Intelligent management of bike sharing in smart cities using machine learning and Internet of Things," *Elsevier,* p. 14, 2020.

[8] P. P. d. Vassimon, "Performance evaluation for bike-sharing systems:," p. 23, 2016.

[9] F. P. F. C. Ji Won Yoon, "Cityride: a predictive bike sharing journey advisor," *IEEE Computer society,* p. 6, 2012.

[10] D. Z. L. W. D. Y. X. M. S. L. Z. W. G. P. T.-M.-T. N. J. J. Longbiao Chen, "Dynamic Cluster-Based Over-Demand Prediction in Bike Sharing Systems," *UBICOMP,* p. 12, 2016.

[11] C. R. J.-B. R. A. E. F. P. F. Pierre Borgnat, "SHARED BICYCLES IN A CITY: A SIGNAL PROCESSING AND DATA ANALYSIS PERSPECTIVE," *World Scientific Publishing Company,* p. 25, 2010.

[12] C. T. C. Z. J. Gallop, "A Seasonal Autoregressive Model Of Vancouver Bicycle Traffic Using Weather Variables," *i-manager's Journal on Civil Engineering,* p. 18, 2011.

[13] "Bixi Montreal," 01 Mar 2019. [Online]. Available: https://www.newswire.ca/news-releases/bixi-montreal-a-record-53-million-rides-in-2018-and-an-expansion-plan-for-five-new-boroughs-700377192.html.

[14] P. DeMaio, "Bike-sharing: History, Impacts,Models of Provision, and Future," Journal of Public Transportation, 2009.