Documentation VLBA II – System Architectures

Topic 7 : Event Based Forecasting

Create a Forecast Model and integrate other data sources to see possible influence

Muralidhar Reddy Kuluru
226240
Magdeburg, 29. Juni 2021

Otto-von-Guericke-Universität Magdeburg
Magdeburg Research and Competence Cluster
Fakultät für Informatik
Institut für Technische und Betriebliche Informationssysteme
Arbeitsgruppe Wirtschaftsinformatik

Table of figures

1. Introduction

To get familiar with Google Cloud Platform and its services, a forecast model is to be built and evaluated. Create a forecasting model for the bike trips in BigQuery. Then try to integrate the weather data to evaluate a possible influence on the overall trips. Describe whether the initial model could be improved this way and try to find additional ideas and data sources to maximize the accuracy of your model.

As per the problem statement , I have to build a forecasting model for Austin_bikeshare dataset for overall biketrips and integrate the weather dataset of Austin region to find any possible influence on the trips using BigQuery, a SaaS offered by Google Cloud Platform.

Tasks to be done:

- Build a Forecast Model using BQML
    - Use biketrips dataset of a region (bigquery-public dataset)
- Integrate with other data source
    - Weather dataset of the region
- Find influence of the weather on overall trips
- Try to find additonal resources that can maximise accuracy

2. Google Cloud Platform (GCP)

Google Cloud Platform

Google Cloud (also known as Google Cloud Platform or GCP) is a provider of computing resources for developing, deploying, and operating applications on the Web. Google Cloud Platform services are robust. One way to navigate them is to consider which solutions are available based on your primary computing needs: infrastructure as a service (IaaS), platform as a service (PaaS), and software-as-a-service (SaaS).

- IaaS enables IT to run virtual machines without having to invest in or manage this computing infrastructure themselves. Often IT will opt for an IaaS solution when the workload is temporary, experimental, or subject to unexpected changes (e.g. sandbox projects).
- PaaS is the next step, building on the IaaS model. Customers opt for all of the benefits of IaaS, plus they get underlying infrastructure – like operating systems and middleware. Their vendor hosts and manages all of these elements.
- SaaS goes one more step – everything is available via the web: the provider hosts, manages, and delivers the entire infrastructure including applications. Users simply log in to access the resources the specific solution delivers, e.g. backup and recovery tools.

Google cloud platform can be used by using console or cloud shell. Major services of Google Cloud include Computing and hosting, Storage and database, Networking, Big Data, Machine learning, Identity and Security and Management tools. Some services offered by Google Cloud Platform are Google Compute Engine(GKE), Google App Engine, Google Kubernetes Engine, Cloud Storage, AutoML, BigQuery and Anthos.

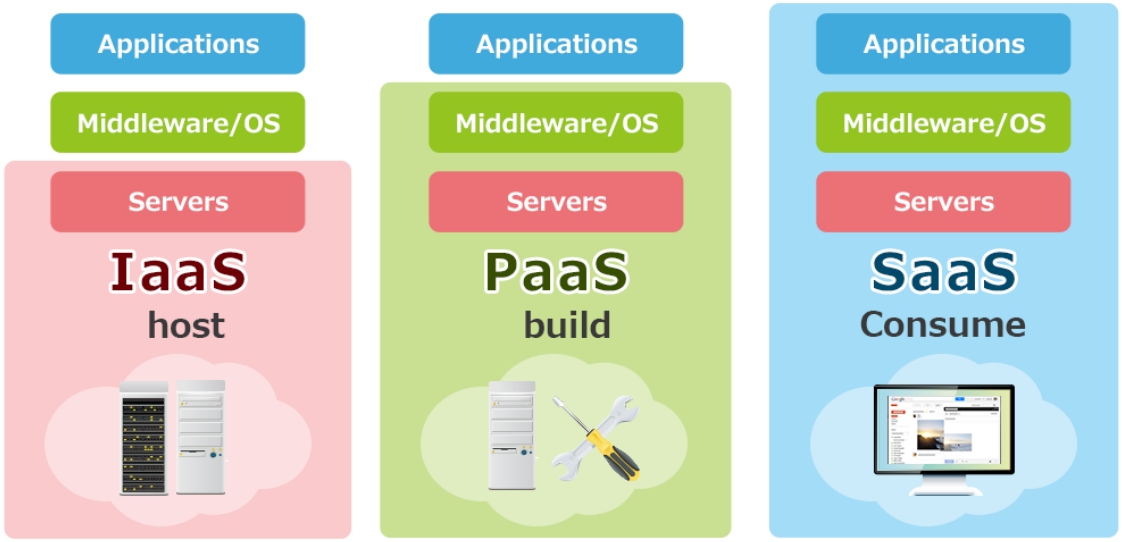


Figure 1 – Cloud Services

3. BigQuery

Storing and querying massive datasets can be time consuming and expensive without the right hardware and infrastructure. BigQuery is an enterprise data warehouse that solves this problem by enabling super-fast SQL queries using the processing power of Google's infrastructure. BigQuery is considered an example of infrastructure as a service (IaaS). It is a serverless, highly scalable data warehouse that comes with a built-in query engine capable of running SQL queries on terabytes of data in a matter of seconds, and petabytes in only minutes. You get this performance without having to manage any infrastructure and without having to create or rebuild indexes. BigQuery can be accesses by using the Cloud Console, by using the bq command-line tool, or by making calls to the BigQuery REST API using a variety of client libraries such as Java, .NET, or Python. There are also a variety of third-party tools that you can use to interact with BigQuery, such as visualizing the data or loading the data.

BQML: BigQuery Machine Learning allows to load data easily, processes model quickly and runs queries in SQL, a language that every data analyst know and easy to learn.

Data Studio: The results can be visualised through Google Data Studio which is interactive and provides lot of options to display the results better.

Advantages:

- Real-time analytics

- Logical data warehousing

- Data transfer services

- Automatic backup and easy restore

- Serverless insight

Fig 2: BigQuery

4. Event Based Forecasting

Event Based Forecasting is a periodic forecasting with a new, event-based process that takes place when something unexpected happens. This forecasting depends on historical data and businesses use forecasting to manage on how to allocate their budgets or plan for anticipated expenses for an upcoming period of time.

Event based planning is not just a phenomenon of a specific company or field, but every business needs to plan and re-forecast based on events. Some events may impact directly such as the company example and others could be indirect impacts such as, competitor action affecting sales revenue, merger and acquisition affecting internal plans etc.

The forecasting model should be carefully designed to deliver comprehensible results.

Linear Regression:

Machine learning, more specifically the field of predictive modeling is primarily concerned with minimizing the error of a model or making the most accurate predictions possible, at the expense of explainability. Linear regression is a basic and commonly used type of predictive analysis. It tries to develop a relationship between two variables by fitting a linear equation to fitting data. These relations are developed using linear predictor functions whose unknown parameters are derived from the data. Linear regression models can be used to fit a predictive model if the goal is prediction, forecasting or in some cases error reduction.

I have used linear regression model to predict/forecast the overall trips made in a year with and without weather data.

5. Datasets and Pre-Processing

Dataset1: Bikeshare_dataset

The first task, developing a forecast model for overall trips has to be developed by using the Austin bikeshare trips dataset. This dataset is publicly available in BigQuery. BigQuery has several datasets that could be accessed for free with or without importing to the GCP users bucket. The dataset was accessed as `*bigquery-public-dataset.austin_bikeshare_bikeshare_trips*`. The dataset austin_bikeshare contains two tables ,*bikeshare_stations* and *bikeshare_trips*.

bikeshare_stations table contains information about the bike stations present in the Austin along with the address, station_id, station_name and few other details. There are 98 bike stations in the city as per the dataset. Second table *bikeshare_trips* contains data about the trips made by several users with columns as trip_id, subscriber_type, bikeid, start_time, start_stataion_id,

start_station_name, end_station_id, end_station_name, duration_minutes. The table contains 9 columns and 1,342,066 rows. From this table, count of trip_id is considered as overall trips and as a label for the regression model. This table contain records from Dec 2013 to Feb 2021.

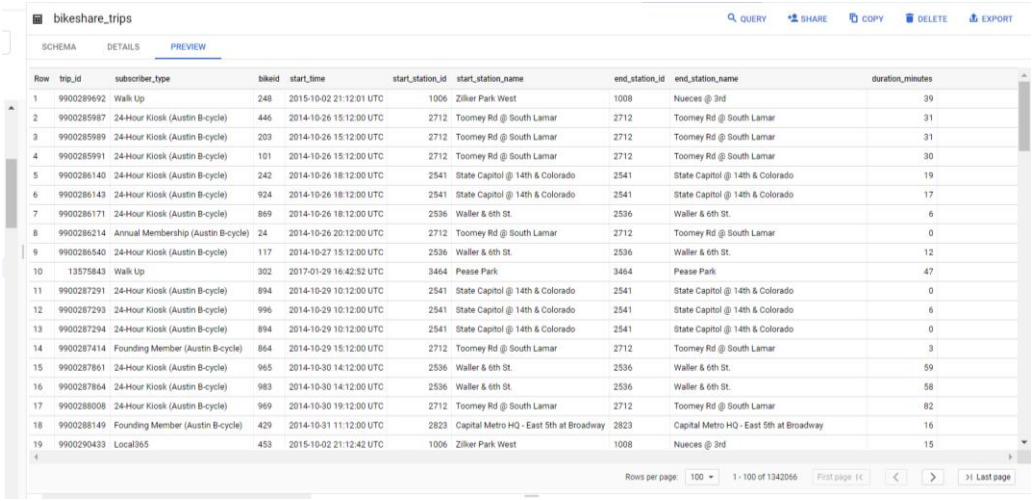
I used only the second table *bikeshare_trips* for building the model as the first table doesn't seem to contain any information regarding trips.

Pre-Processing:

NULL Values : The dataset has null values only in start_station_id column(23154) but we have start_station_name as another column that contain names of the start_stations and so, these missing values are not an issue. These null values constitutes only 1 percent of the dataset and can be omitted.

Conversion:

- Type Conversion: The values of bikeid and end_station_id columns were in String type and were converted to integer type.
- Start_time: Values of the column start_time were in timestamp type and format 'YYYY-MM-DD HH:MM:SS UTC'. Day, Month and Hour values were extracted from the dataset into separate columns.



Row	trip_id	subscriber_type	bikeid	start_time	start_station_id	start_station_name	end_station_id	end_station_name	duration_minutes
1	9900289692	Walk Up	248	2015-10-02 21:12:01 UTC	1006	Zilker Park West	1008	Nueces @ 3rd	39
2	9900285987	24-Hour Kiosk (Austin B-cycle)	446	2014-10-26 15:12:00 UTC	2712	Toomey Rd @ South Lamar	2712	Toomey Rd @ South Lamar	31
3	9900285989	24-Hour Kiosk (Austin B-cycle)	203	2014-10-26 15:12:00 UTC	2712	Toomey Rd @ South Lamar	2712	Toomey Rd @ South Lamar	31
4	9900285991	24-Hour Kiosk (Austin B-cycle)	101	2014-10-26 15:12:00 UTC	2712	Toomey Rd @ South Lamar	2712	Toomey Rd @ South Lamar	30
5	9900286140	24-Hour Kiosk (Austin B-cycle)	242	2014-10-26 18:12:00 UTC	2541	State Capitol @ 14th & Colorado	2541	State Capitol @ 14th & Colorado	19
6	9900286143	24-Hour Kiosk (Austin B-cycle)	924	2014-10-26 18:12:00 UTC	2541	State Capitol @ 14th & Colorado	2541	State Capitol @ 14th & Colorado	17
7	9900286171	24-Hour Kiosk (Austin B-cycle)	869	2014-10-26 18:12:00 UTC	2536	Waller & 6th St.	2536	Waller & 6th St.	6
8	9900286214	Annual Membership (Austin B-cycle)	24	2014-10-26 20:12:00 UTC	2712	Toomey Rd @ South Lamar	2712	Toomey Rd @ South Lamar	0
9	9900286540	24-Hour Kiosk (Austin B-cycle)	117	2014-10-27 15:12:00 UTC	2536	Waller & 6th St.	2536	Waller & 6th St.	12
10	13575843	Walk Up	302	2017-01-29 16:42:52 UTC	3464	Pease Park	3464	Pease Park	47
11	9900287291	24-Hour Kiosk (Austin B-cycle)	894	2014-10-29 10:12:00 UTC	2541	State Capitol @ 14th & Colorado	2541	State Capitol @ 14th & Colorado	0
12	9900287293	24-Hour Kiosk (Austin B-cycle)	996	2014-10-29 10:12:00 UTC	2541	State Capitol @ 14th & Colorado	2541	State Capitol @ 14th & Colorado	6
13	9900287294	24-Hour Kiosk (Austin B-cycle)	894	2014-10-29 10:12:00 UTC	2541	State Capitol @ 14th & Colorado	2541	State Capitol @ 14th & Colorado	0
14	9900287414	Founding Member (Austin B-cycle)	864	2014-10-29 15:12:00 UTC	2712	Toomey Rd @ South Lamar	2712	Toomey Rd @ South Lamar	3
15	9900287861	24-Hour Kiosk (Austin B-cycle)	965	2014-10-30 14:12:00 UTC	2536	Waller & 6th St.	2536	Waller & 6th St.	59
16	9900287864	24-Hour Kiosk (Austin B-cycle)	983	2014-10-30 14:12:00 UTC	2536	Waller & 6th St.	2536	Waller & 6th St.	58
17	9900288008	24-Hour Kiosk (Austin B-cycle)	969	2014-10-30 19:12:00 UTC	2712	Toomey Rd @ South Lamar	2712	Toomey Rd @ South Lamar	82
18	9900288149	Founding Member (Austin B-cycle)	429	2014-10-31 11:12:00 UTC	2823	Capital Metro HQ - East 5th at Broadway	2823	Capital Metro HQ - East 5th at Broadway	16
19	9900290433	Local365	453	2015-10-02 21:12:42 UTC	1006	Zilker Park West	1008	Nueces @ 3rd	15

Fig 3: Biketrips_Dataset

Dataset2: Weather Data

This dataset was downloaded from Kaggle and the purpose of the dataset it to estimate the influence of weather on overall trips of the model. This dataset contain weather data in Austin region from Dec 2013 to July 2017. Weather data includes Temperature, Humidity, Dew point, Visibility, Wind speed, Sea level pressure, precipitation and events. The dataset has total of 1319 records in total.

A new dataset was created in BigQuery for this project. A Cloud storage bucket was created and the downloaded dataset was loaded into the bucket as a one-time load. Under the dataset, a table was created using the data loaded into cloud storage.

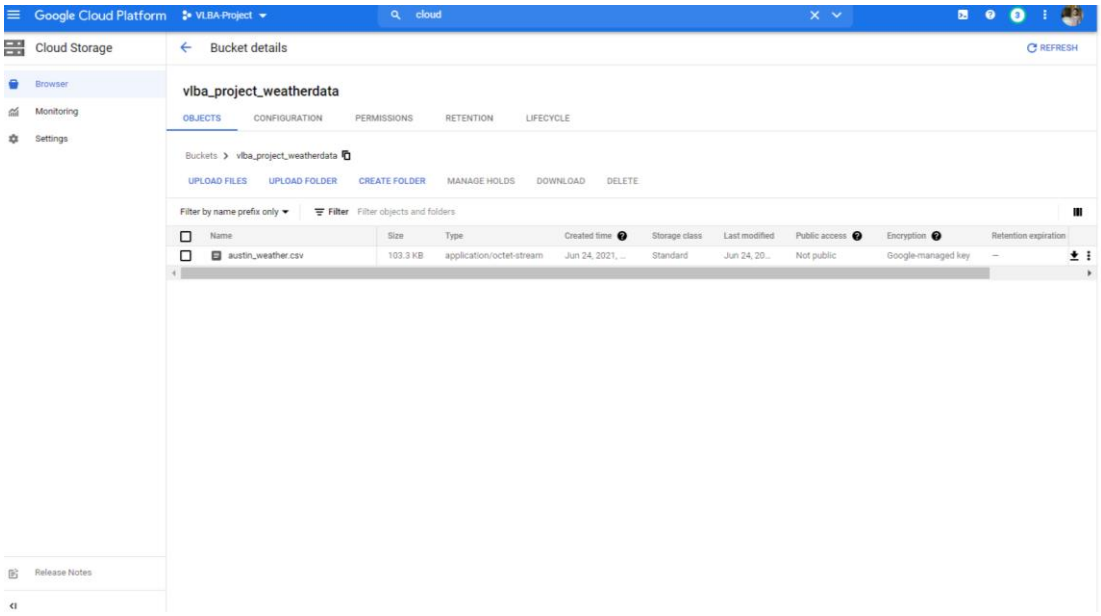


Fig 4: Cloud Storage

Pre-Processing:

The weather data had null values in SeaLevelPressure and this column was omitted while building the model. All the columns except temperature were in string format and were converted to integer values.

The weather dataset needs to be integrated with the bikeshare_trips and a left join was performed to combine both the datasets. The newly formed dataset was stored as a table in the created dataset. The new dataset had around 40% of missing values as the trips data was from 2013 to 2021 and weather data is only from 2013 t0 2017. There is missing data for almost four years and this cannot be removed because removing this will result in lots of data.

Missing Data: The resulted dataset had almost 40% of rows without any weather values and these cannot be removed. The must be imputed with any possible way available. I have tried to impute data in three ways and I have used the last way finally and stored the data.

- *Mean* : Mean of the columns i.e, calculate mean of the temperature from 2013 to 2017 and replace the missing values that is from ’17 to ’21 with the value. This wasn’t a feasible was as the values for lot of data would be same and the model won’t be efficient.
- *Random Value*: Take any random value in the range of min and max value of a column and replace the null values with it. This method randomly chose a value and in some cases assigned different temperatures to different months and didn’t seem efficient.
- *Mean value based on the months*: Calculate the mean value of particular month and replace the null values of the corresponding month with that data. This resulted in having same weather for a month in all three years.

I have imputed the null values with the mean value based on the months and built the model.

	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R
	bikeid	start_time	end_station_id	duration_minute	TempHighF	TempAvgF	HumidityAvgPercent	HumidityLowPercent	VisibilityHighMile	VisibilityAvgMile	VisibilityLowMile	WindHighMPH	WindHighMPH_	WindHighMPH_2	
1	-0.1855	0.0094	-0.3983	-0.0104	-0.0877	-0.0845	-0.0257	-0.0076	-0.0094	-0.0179	-0.0041	0.0047	0.0047	0.0047	
5	1	0.0262	0.1503	0.0194	0.0041	0.0046	0.0074	0.0054	0.0037	-0.0001	-0.0012	-0.0055	-0.0055	-0.0055	
4	0.0262	1	0.0211	-0.0085	0.0162	0.0146	-0.0082	-0.0094	-0.0011	0.0038	0.001	-0.0122	-0.0122	-0.0122	
3	0.1503	0.0211	1	-0.004	0.0386	0.0391	0.0144	0.0027	0.0002	0.0124	0.0001	0.019	0.019	0.019	
4	0.0194	-0.0085	-0.004	1	0.0024	0.0027	0.0035	0.0044	0.0021	-0.0036	-0.0035	0.0019	0.0019	0.0019	
7	0.0041	0.0162	0.0386	0.0024	1	0.9627	-0.093	-0.2169	0.0727	0.208	0.1523	0.041	0.041	0.041	
5	0.0046	0.0146	0.0391	0.0027	0.9627	1	0.0695	-0.0283	0.0633	0.1218	0.0573	0.0491	0.0491	0.0491	
3	0.005	0.0125	0.0374	0.0028	0.8628	0.9658	0.2211	0.1544	0.0495	0.0285	-0.04	0.0521	0.0521	0.0521	
1	0.0078	0.0082	0.0418	0.0041	0.7677	0.8613	0.4693	0.3457	0.031	-0.1402	-0.2409	0.1305	0.1305	0.1305	
2	0.0084	0.0087	0.0422	0.0045	0.7659	0.8736	0.5214	0.4101	0.0294	-0.1363	-0.2274	0.0635	0.0635	0.0635	
9	0.0087	0.0116	0.0387	0.0042	0.7166	0.833	0.5295	0.4552	0.0266	-0.0942	-0.1811	-0.0142	-0.0142	-0.0142	
3	0.0071	-0.0039	0.0219	0.0013	0.1125	0.1906	0.8249	0.5389	-0.0309	-0.4133	-0.5082	0.0159	0.0159	0.0159	
7	0.0074	-0.0082	0.0144	0.0035	-0.093	0.0695	1	0.9161	-0.0689	-0.522	-0.5878	0.0164	0.0164	0.0164	
5	0.0054	-0.0094	0.0027	0.0044	-0.2169	-0.0283	0.9161	1	-0.0824	-0.4878	-0.5226	0.0088	0.0088	0.0088	
4	0.0037	-0.0011	0.0002	0.0021	0.0727	0.0633	-0.0689	-0.0824	1	0.154	0.0539	0.0403	0.0403	0.0403	
9	-0.0001	0.0038	0.0124	-0.0036	0.208	0.1218	-0.522	-0.4878	0.154	1	0.8175	-0.0915	-0.0915	-0.0915	
1	-0.0012	0.001	0.0001	-0.0035	0.1523	0.0573	-0.5878	-0.5226	0.0539	0.8175	1	-0.1763	-0.1763	-0.1763	
7	-0.0055	-0.0122	0.019	0.0019	0.041	0.0491	0.0164	0.0088	0.0403	-0.0915	-0.1763	1	1	1	
7	-0.0055	-0.0122	0.019	0.0019	0.041	0.0491	0.0164	0.0088	0.0403	-0.0915	-0.1763	1	1	1	
7	-0.0055	-0.0122	0.019	0.0019	0.041	0.0491	0.0164	0.0088	0.0403	-0.0915	-0.1763	1	1	1	

Fig 5: Corelation between variables

Conversion: All the columns except for the weather were in String format and they were converted into Integer format.

6. Forecast Model

The first task is to build a forecast model to estimate the overall trips in BigQuery by using bikeshare_trips table. I tried to find the correlation between the features to know the relation between the features and how they are affected. But the correlation wasn't a good choice for this and the results didn't come as expected. By running few SQL queries, I tried to figured out what factors affect the data most and have visulalized the effect of few features on the overall trips.

Training Data : Trips before start date'2018-01-01'

Test Date : Trips on and after '2018-01-01'

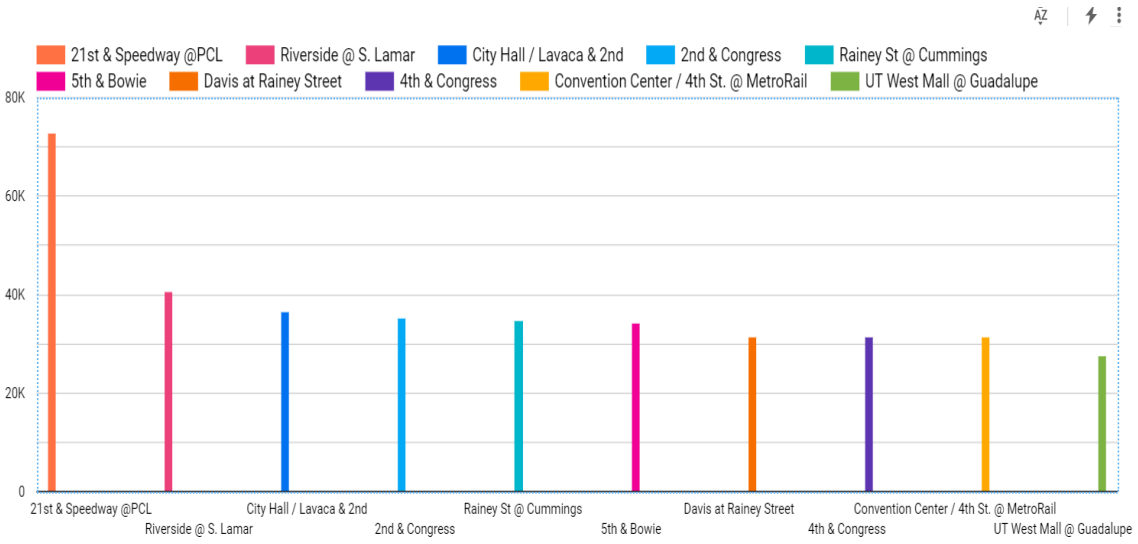


Fig 6: Stations with most trips

Start_station_name '21st&Speedway @PCL` turned out to have more number of trips and with station id '3798'.

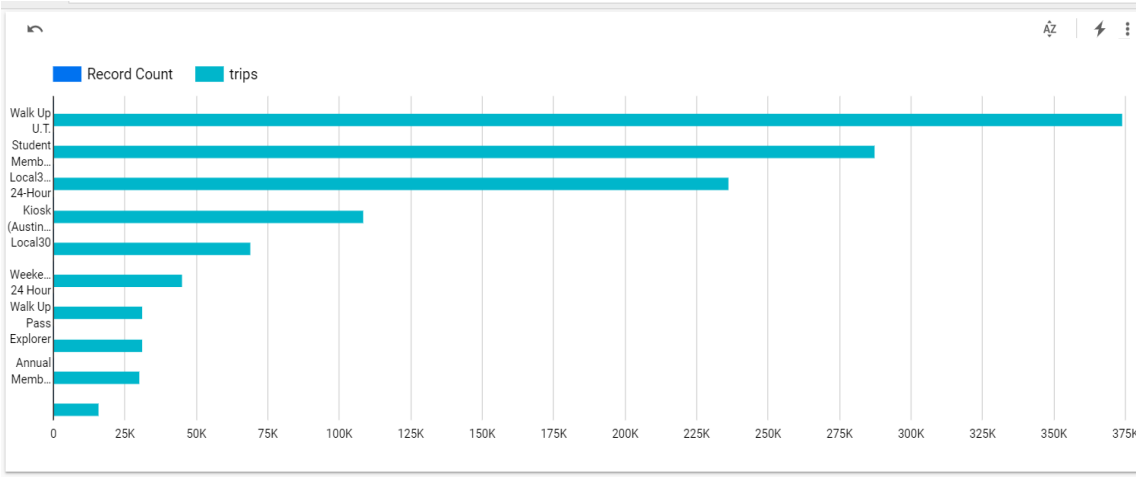


Fig 7: Trips based on Subscriber type

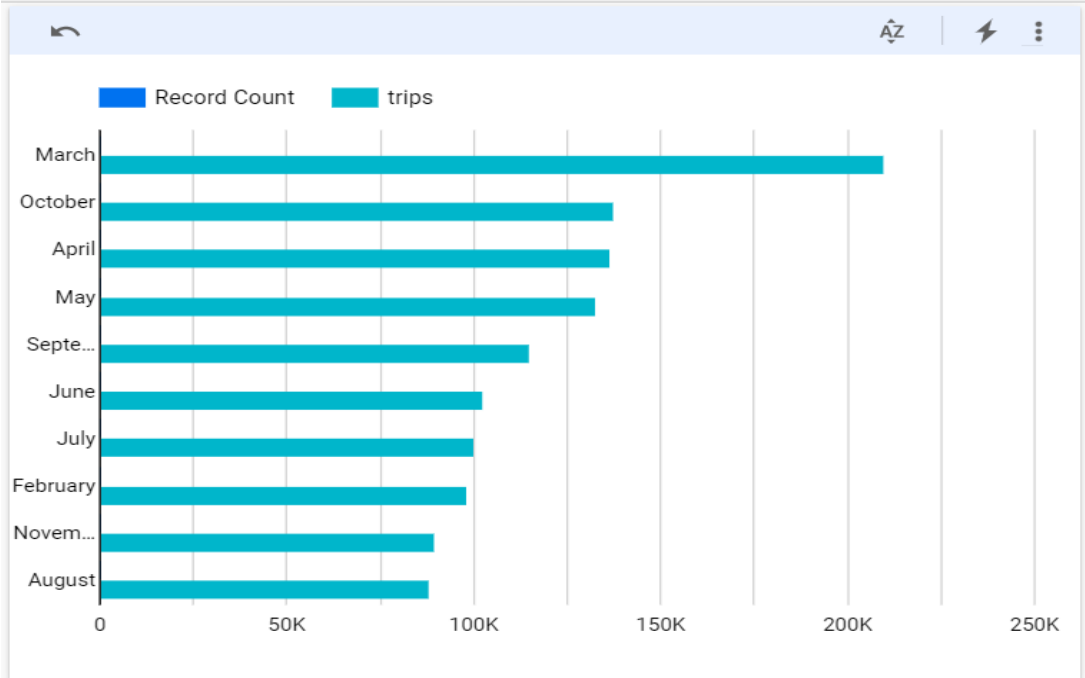


Fig 8: Trips wrt to Months(Training Data)

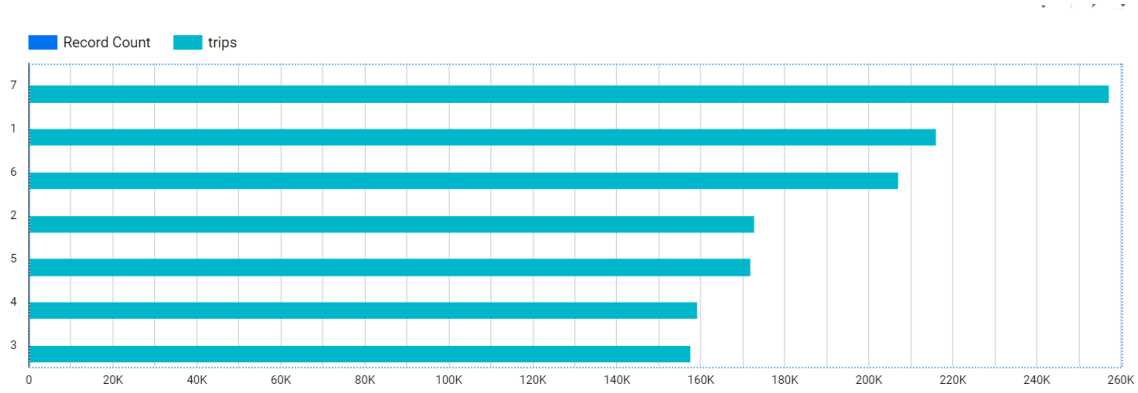


Fig 9: Trips wrt to Day of Week

Based on the above graphs, it is inferred that the more number of trips oocured with station '21st &Speedway @PCL' and with subscriber id 'Walk Up'. More number of trips occurred on Saturday and Sunday being weekends. I used these features foe building the model and then predicted the overall trips.

Building Model:

I have considered the features and have tried to implement basic linear regression model with data before '2018-0-01' as training data and later did some hypertuning on the parameters to estimate the overall trips. The model resulted with mean_absolute_error of 24 and r2_square of 0.14.

Features:

- Start_station_name
- Start_station_id
- Subscriber_type
- Start_day
- Start_month and start_hour
- Duration_Minutes

Below is the snippet of the training summary of the model:

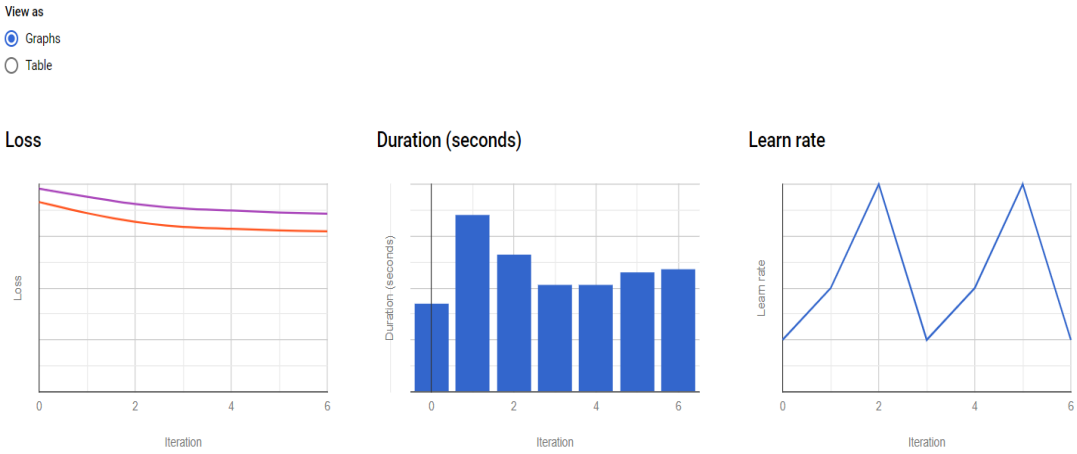


Fig 10: Summary of biketrips_model

Fig 11: Evaluation biketrips_model

Parameters:

Optimise_strategy = Batch Gradient Descent

Data_Split = Auto_Split

Max_Iterations = 20

Learn_Rate_strategy = LineSearch

7. Model with Weather Data

The above model was developed considering the bikeshare_trips data only and this model will be developws with the table that contains both the trips data and weather data. For building this model, I have considered the same features as above from the trips table and from weather table average temperature, average humidity, average dew point, average visibility, average wind speed were considered. Only the averages were considered because the high and low values of the weather didn't seem to be as efficient features and the number of features were increasing.

Features:

- Start_station_name
- Start_station_id
- Subscriber_type
- Start_day
- Start_month and start_hour
- Duration_Minutes
- TempAvgF
- HumidityAvgPercent
- DewPointAvg
- VisibilityAvgMiles
- WindSpeedMPH

Below is the snippet of the model summary:

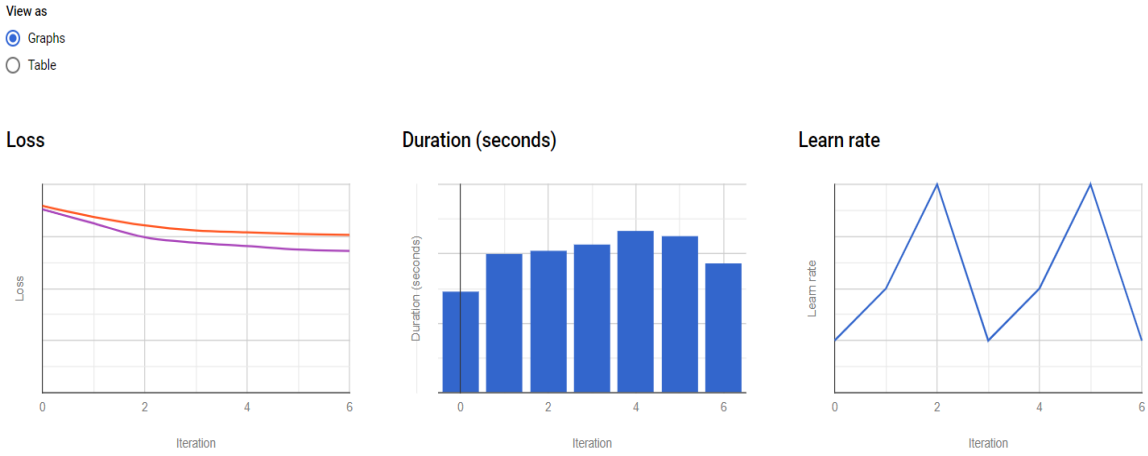


Fig 12: Summary of trips_weather model

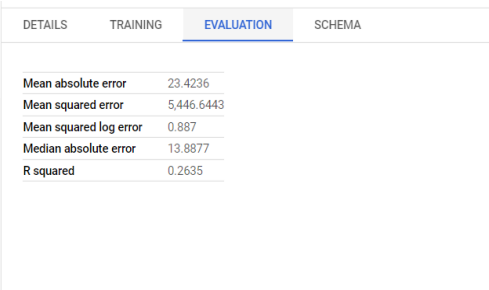


Fig 13: Evaluation Summary of trips_weather model

Parameters:

Optimise_strategy = Batch Gradient Descent

Data_Split = Auto_Split

Max_Iterations = 20

Learn_Rate_strategy = LineSearch

8. Evaluation and Results

After building the above models, I evaluated them with the data from '2018-01-01' and haven´t seen much of a difference in the evaluation error. Even after trying with different features and tuning model parameters, the change doesn't seem to be good and wasn't affected much.

I tried to predict the total number of trips made after 2019 for the station_id 3798 which is the station id for '21st &Speedway@PCl' and '21st /Speedway@PCL' and there is a slight change in the number of average trips and daily average trips made.

Job information | Results | JSON | Execution details

Row	start_station_name	predicted_total_rides	avg_predicted_rides_daily	actual_total_rides	actual_avg_rides_daily
1	21st/Speedway @ PCL	4747.365166477084	13.006479908156395	7381	20.221917808219178
2	21st & Speedway @PCL	4290.4266375115185	11.754593527428817	6662	18.252054794520546

Fig:14 Predictions made only with trips

The model built only with bikeshare_trips datset predicted the total rides as 4747 and 4290.

Job information | Results | JSON

Row	start_station_name	predicted_total_rides	avg_predicted_rides_daily	actual_total_rides	actual_avg_rides_daily
1	21st & Speedway @PCL	5041.723761569676	13.812941812519659	6662	18.252054794520546
2	21st/Speedway @ PCL	5403.956284713309	14.805359684146053	7381	20.221917808219178

The model built with weather data combined predicted 5403 and 5041 out of 7381 and 6662 rides. The model I built wasn't that accurate but it shows that the number of rides have increased when weather data was also added to the original dataset.

Additional Resources:

I have tried to find additional datasets that could provide the weather values in austin from 2017 to 2021 but unfortunately, I didn't get any datasets with desired data. There was one dataset that has only temperature values from 2013-2018 Feb and integrating that as well didn't give an fruitful results. I couldn't find any datasets in Kaggle and also any sources from where I could source the data.

9. Conclusion

I have developed the model for overall trips with the predicted data and integrated it with the weather data and developed an another model to see the impact on the overall performance. From the data, it is inferred that there is lot of null values that has been imputed with the mean values based on the values of previous months. The models built have resulted in with a high mean_absolute_error. The predictions were made based on the station with more trips where the total rides and average rides per day were predicted. The results shows that there is a slight variation in the prediction when weather data is also included in the model.

The model could be further improved if the missing values were imputed in a better way and if the missing weather data is found. I have tried to find any additional resources that could be helpful in increasing the accuracy but couldn't find a better data source.

10. References

[1] https://cloud.google.com/bigquery/docs

[2] https://www.kaggle.com/grubenm/austin-weather

[3] https://www.zdnet.com/article/what-is-google-cloud-is-and-why-would-you-choose-it/

[4] https://cloud.google.com/docs/

[5] https://cloud.google.com/blog/products/gcp