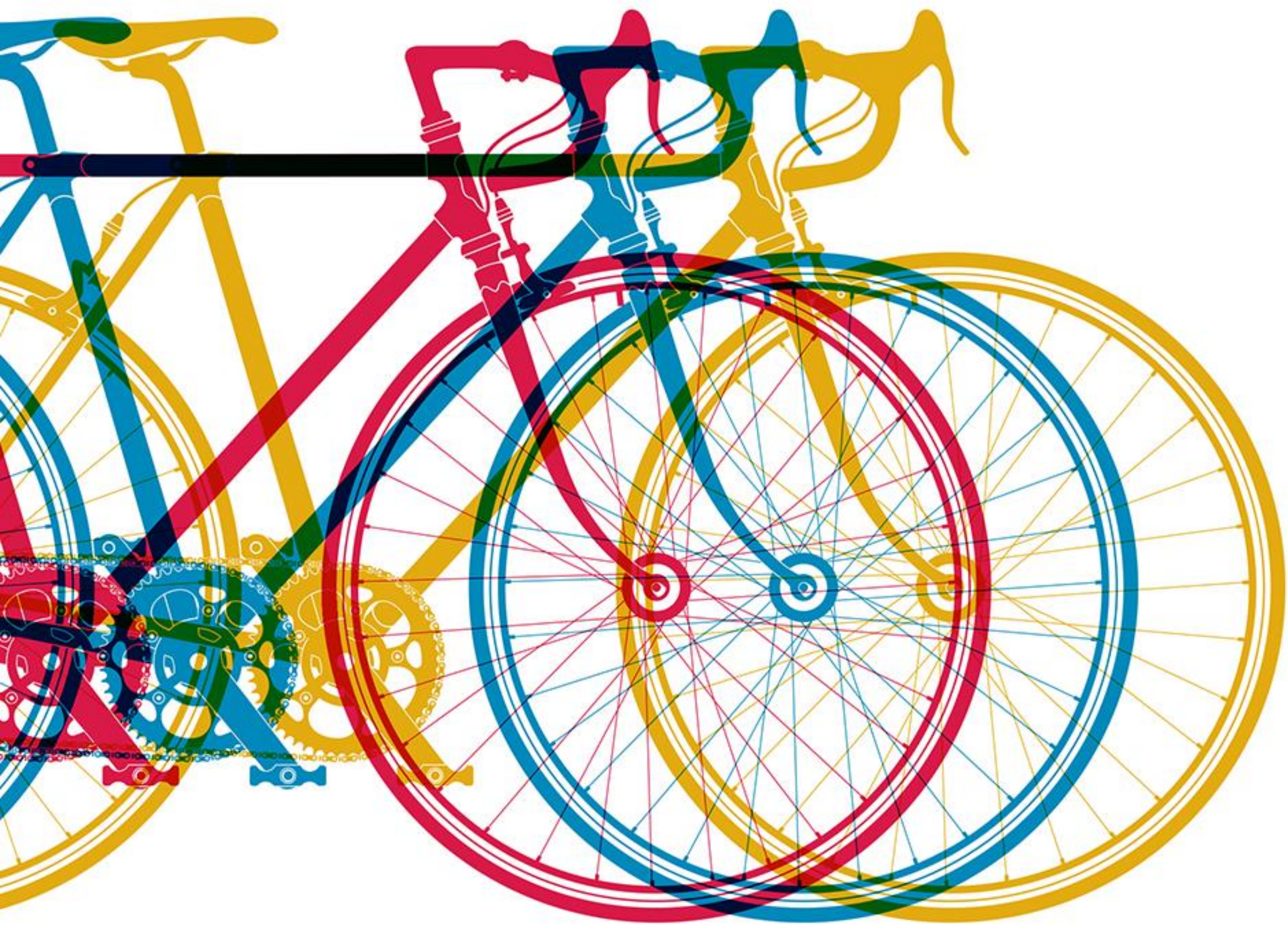
PREDICTING SHARED BIKE RENTALS



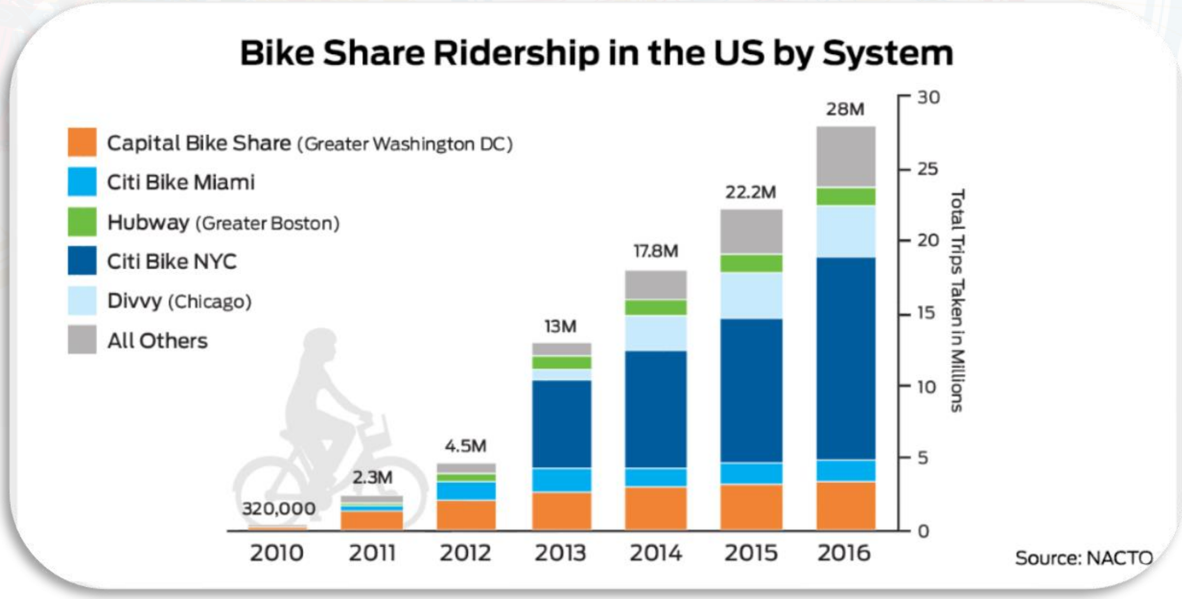
Project By:
Nitansh Gupta
NXG180004

Index

Abstract

Climate change and pollution has been very big concern all over the globe. While the list of factors causing this change is huge, ever increasing presence of motor vehicles has persisted as one of the major demonic source of the air pollution. While electric vehicles are pitch for a lucrative potential solution, their reach and and quality is still not at par. Within the list of high potential solutions, Bike Sharing has appeared as one of the most successful alternative, having shown success in many parts of the world already. While even the less developed nations across Asia, and Europe have implemented the bike sharing model with much established infrastructure, US still has a lot more potential untapped.

Bike share is growing at an astounding clip across the U.S., with over **88 million** trips made on a bike share bike in the U.S. since 2010. In 2016 alone, riders took over **28 million** trips



This could be because of many basic reasons like availability, ease-of-use, quality of bikes and related services, mindset of the populace, promotional activities etc. Being at a nascent stage, these variables affecting the adoption of bike sharing ecosystem can not be quantified. So if one wants to enquire and predict the use of the shared bikes, macro factors, like weather and event types could be used most effectively to initiate this developmental phase.

Approach

We got the data of 2 years of shared bike rentals through rentals company named CAPITAL BIKESHARE, which is metro DC's bikeshare service, with 4,300 bikes and 500+ station across 7 jurisdictions. The bikes renters are of two types: Registered – renters who have taken periodic rental subscriptions and hence have more re-occurring use. Casual renters on the other hand are the once a while users who rent out on one time payment basis. These casual and registered rentals are mapped against the macro factors like temperature, humidity, windspeed , working-day/holiday(flags) etc. We not only looked at the trends but also evaluated various supervised learning models to predict each kind of rentals. In the process, we executed data cleaning, EDA, feature engineering, model execution and evaluation.

About Data

Source: https://archive.ics.uci.edu/ml/datasets/Bike+Sharing+Dataset

Contributor: Capital BikeShare, Washington DC

Shape: 17,389 x 17

Hourly rentals for the metro DC area over span of two years

The Temperature, Humidity,and Windspeed are provided after normalization

Index
- Instance_id
- rental_date

Target Variable
- Total_rentals

Continuous Variables
- temp
- temp_feel
- humidity
- windspeed
- casual_rental
- registered_rentals

Categorical Variables
- season
- Is_holiday
- weather_condition
- month
- year
- hour
- is_workingday
- weekday

season

1:winter
2:spring
3:summer
4:fall

Weather

1: Clear, Few clouds, Partly cloudy, Partly cloudy

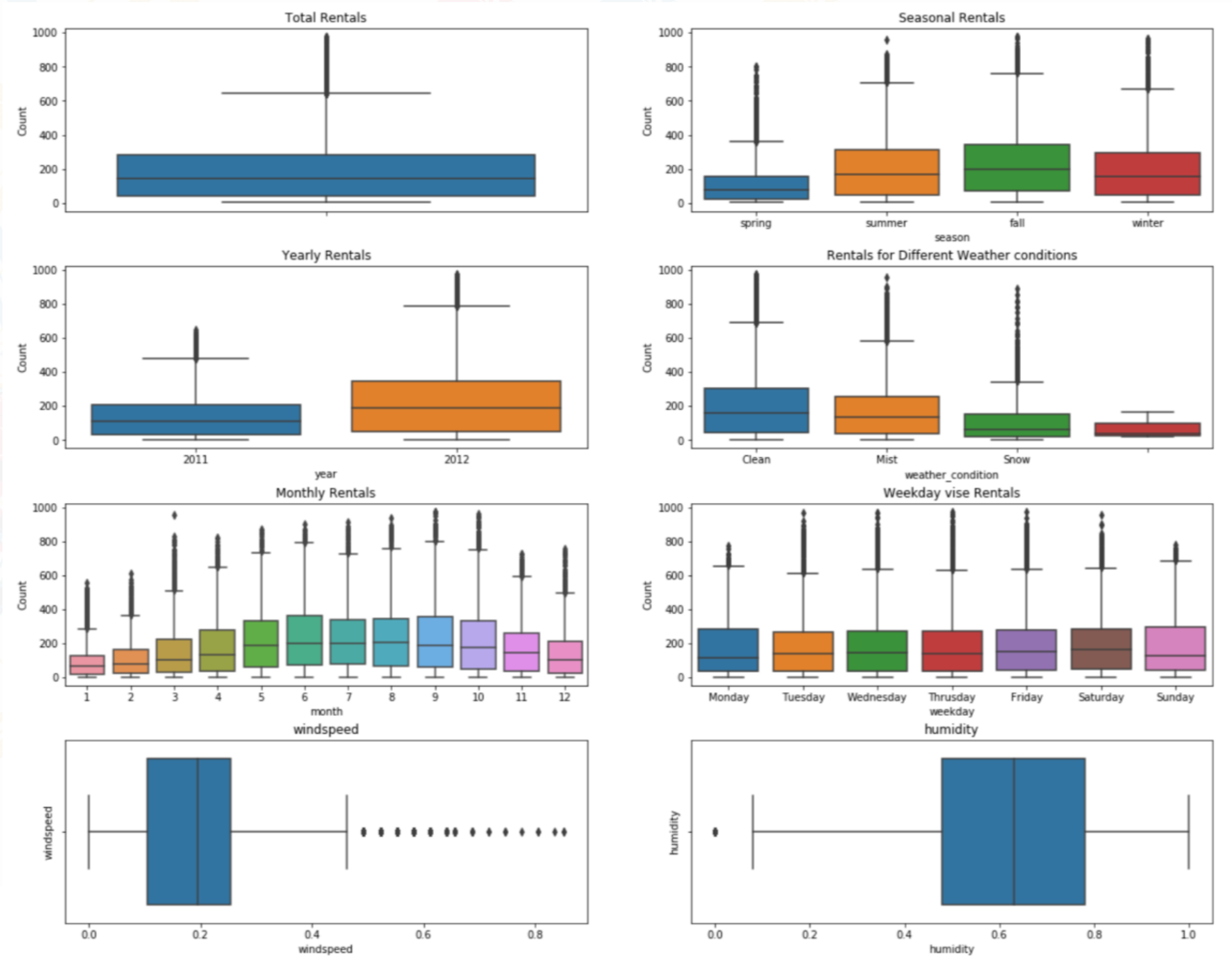
2: Mist + Cloudy, Mist + Broken clouds, Mist + Few clouds, Mist

3: Light Snow, Light Rain + Thunderstorm + Scattered clouds, Light Rain + Scattered clouds

4: Heavy Rain + Ice Pallets + Thunderstorm + Mist, Snow + Fog

	instance_id	temp	temp_feel	humidity	windspeed	casual_rentals	registered_rentals	total_rentals
count	17379.0000	17379.000000	17379.000000	17379.000000	17379.000000	17379.000000	17379.000000	17379.000000
mean	8690.0000	0.496987	0.475775	0.627229	0.190098	35.676218	153.786869	189.463088
std	5017.0295	0.192556	0.171850	0.192930	0.122340	49.305030	151.357286	181.387599
min	1.0000	0.020000	0.000000	0.000000	0.000000	0.000000	0.000000	1.000000
25%	4345.5000	0.340000	0.333300	0.480000	0.104500	4.000000	34.000000	40.000000
50%	8690.0000	0.500000	0.484800	0.630000	0.194000	17.000000	115.000000	142.000000
75%	13034.5000	0.660000	0.621200	0.780000	0.253700	48.000000	220.000000	281.000000
max	17379.0000	1.000000	1.000000	1.000000	0.850700	367.000000	886.000000	977.000000

Data Distribution

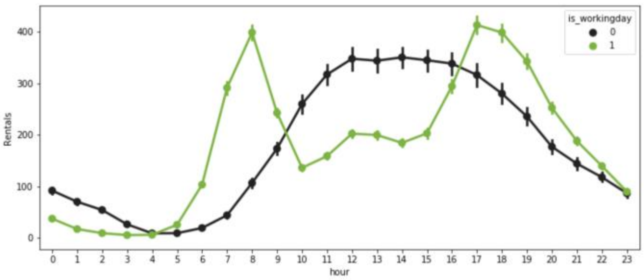


Correlations

	instance_id	temp	temp_feel	humidity	windspeed	casual_rentals	registered_rentals	total_rentals
instance_id	1	0.118635	0.120389	0.0142912	-0.0734779	0.125453	0.228985	0.223319
temp	0.118635	1	0.988454	-0.0660389	-0.0112484	0.462452	0.323255	0.399863
temp_feel	0.120389	0.988454	1	-0.0498397	-0.0497124	0.456303	0.321468	0.396587
humidity	0.0142912	-0.0660389	-0.0498397	1	-0.271258	-0.344293	-0.290944	-0.338566
windspeed	-0.0734779	-0.0112484	-0.0497124	-0.271258	1	0.101616	0.103371	0.11415
casual_rentals	0.125453	0.462452	0.456303	-0.344293	0.101616	1	0.521623	0.720482
registered_rentals	0.228985	0.323255	0.321468	-0.290944	0.103371	0.521623	1	0.967475
total_rentals	0.223319	0.399863	0.396587	-0.338566	0.11415	0.720482	0.967475	1

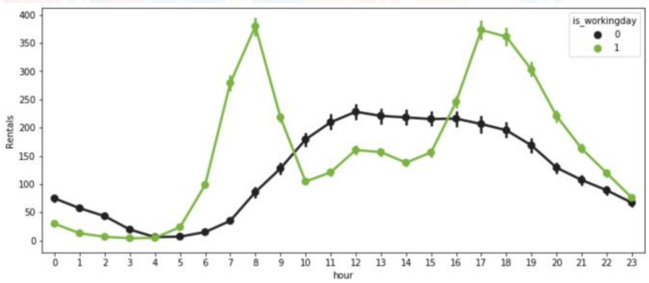
Time Trends

Total Rentals Hourly Trends

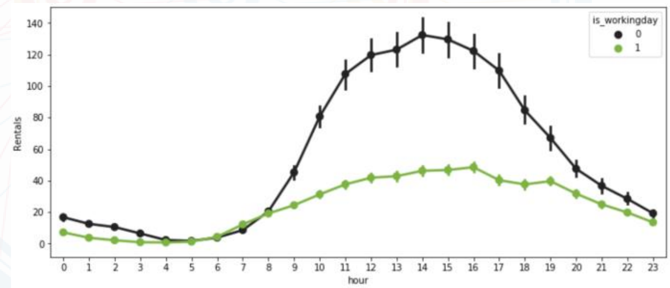


When looking at hourly average of rentals, the casual renters have similar pattern, on working and non working days. Where as for registered rentals the peaks can be seen at the office start and end timings

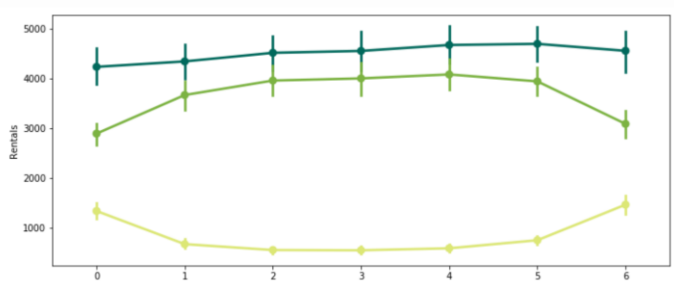
Registered Rentals Hourly Trends



Casual Rentals Hourly Trends

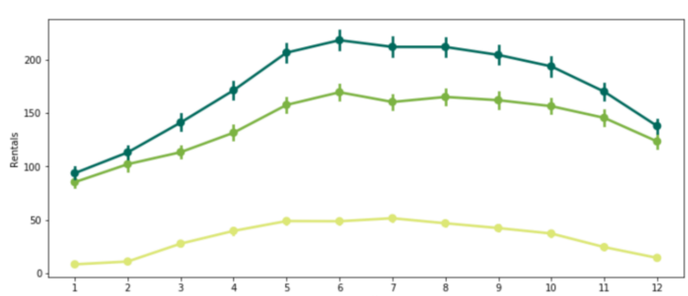


Rentals Trends by Day of Week



Casual rentals and registered rentals show mirror image trends when their average day of week counts are compared. One must place 80% of campaigns for registered renters during the mid of week

Rentals Trend by Month

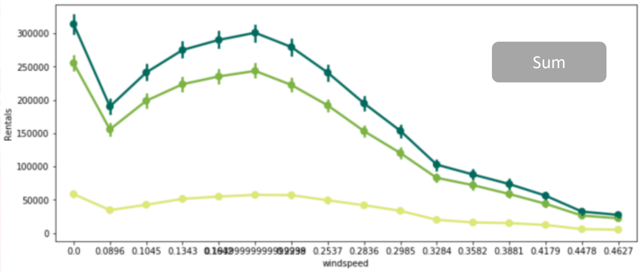


There is a slight dip from June to July for registered rentals, where as there is negligible uplift for casual rentals. We can guess that the increasing temperature during those months might cause the same

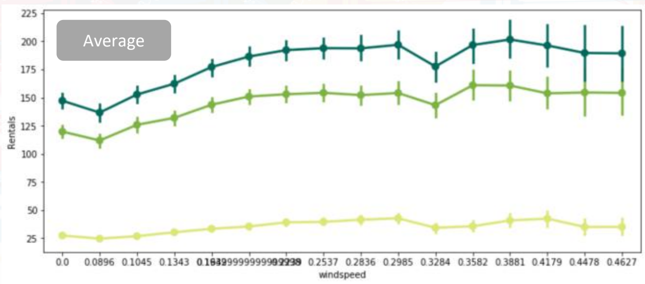
--Total_Rentals --Registered_Rentals --Casual_Rentals

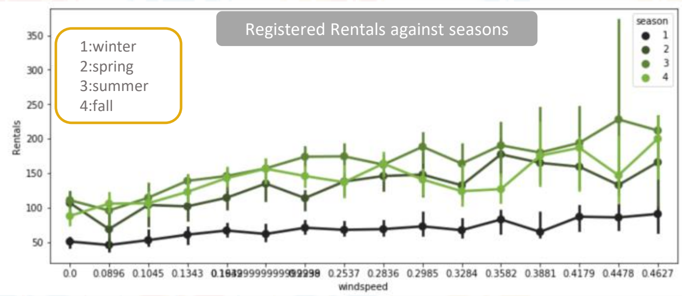
Seasonality Trends

Windspeed vs Rentals

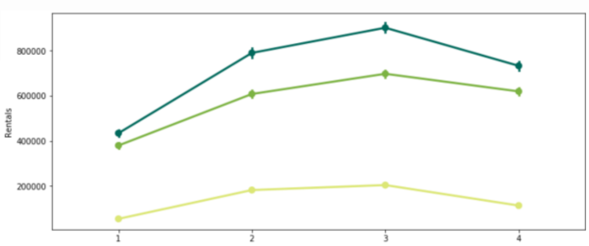


We looked at the trends of rentals against wind speed. While in totality the counts are decreasing, on an average the count was increasing. This meant the there is high variance in the counts number for rentals during high wind speed. On further investigation, we found through the below graph that this variance is caused during the summer season, when high wind may not mean bad weather to ride





Season vs Rentals

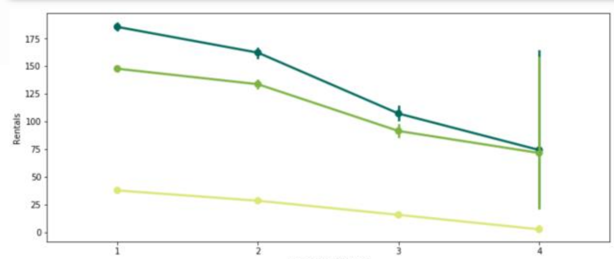


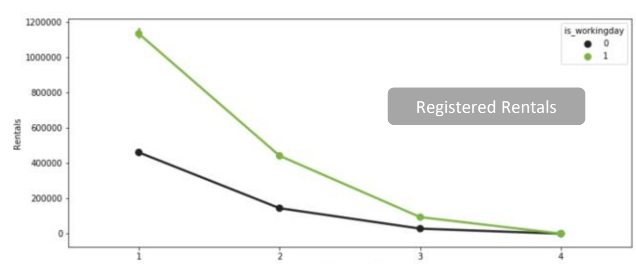
Summer and Spring are the favorite seasons for the bike renters.

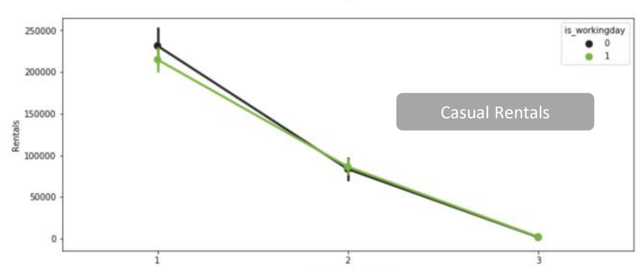
Rentals on working days see more drastic dip in rentals with worsening of weather. Where as non working day rentals and casual rentals don't have that significant drop unless the conditions are extreme

1: Clear, Few clouds, Partly cloudy, Partly cloudy
2: Mist + Cloudy, Mist + Broken clouds, Mist + Few clouds, Mist
3: Light Snow, Light Rain + Thunderstorm + Scattered clouds, Light Rain + Scattered clouds
4: Heavy Rain + Ice Pallets + Thunderstorm + Mist, Snow + Fog

Weather vs Rentals







Applying Models

As instructed during presentation, we implemented al the models separately for casual and registered rentals, against running models for the combined count

Linear Regression with Cross Validation

Cross Validation Score - Registered			
Train	0.807004	0.822618	0.83958
Test	0.89823	0.840447	0.759312

Cross Validation Score - Casual			
Train	0.653747	0.689019	0.713315
Test	0.741412	0.731359	0.714739

Ridge Regression

Registered Rentals						Casual Rentals				
α = 0.01	α = 0.1	α = 1	α = 10	α = 100		α = 0.01	α = 0.1	α = 1	α = 10	α = 100
0.84303	0.84303	0.84302	0.84236	0.81972	Train	0.73998	0.73998	0.73997	0.73957	0.72386
0.86646	0.86646	0.86643	0.86468	0.82850	Test	0.73110	0.73112	0.73126	0.73208	0.71713

Lasso Regression

Registered Rentals						Casual Rentals				
α = 0.01	α = 0.1	α = 1	α = 10	α = 100		α = 0.01	α = 0.1	α = 1	α = 10	α = 100
0.84303	0.84303	0.84300	0.84041	0.77585	Train	0.73998	0.73997	0.73981	0.73251	0.58533
0.86582	0.86577	0.86519	0.86151	0.77935	Test	0.73207	0.73194	0.72994	0.71434	0.52095

Polynomials Regression

Registered Rentals	n = 1	n = 2
Train	0.84303	0.95714
Test	0.86646	-4.61E+23

Casual Rentals	n = 1	n = 2
Train	0.73998	0.93474
Test	0.73110	-4.20E+23

Linear SVM - Registered Rentals	Linear SVM - Casual Rentals
Best score on C-validation set: 0.79	Best score on C-validation set: 0.67
Best parameters: {'C': 100}	Best parameters: {'C': 100}
Train set score with best parameters: 0.26	Train set score with best parameters: 0.44
Test set score with best parameters: 0.31	Test set score with best parameters: 0.50

RBF SVM - Registered Rentals	RBF SVM - Casual Rentals
Best score on C-validation set: 0.32	Best score on C-validation set: 0.46
Best parameters: {'C': 100, 'gamma': 0.01}	Best parameters: {'C': 100, 'gamma': 0.01}
Train set score with best parameters: 0.24	Train set score with best parameters: 0.43
Test set score with best parameters: 0.26	Test set score with best parameters: 0.45

Poly SVM - Registered Rentals	Poly SVM - Casual Rentals
Best score on C-validation set: 0.21	Best score on C-validation set: 0.31
Best parameters: {'C': 100, 'epsilon': 0.01}	Best parameters: {'C': 100, 'epsilon': 0.01}
Train set score with best parameters: 0.43	Train set score with best parameters: 0.43
Test set score with best parameters: 0.31	Test set score with best parameters: 0.50

Conclusion

We looked at the hourly data of Bike rentals collected for two years. After proper EDA certain data engineering steps were taken which lead to average to good test-train scores for all the. Amongst all the models **Lasso Regression** was the best. When fitted on the whole dataset it gave score of 0.85252 for alpha = 0.1

Going further with quest to have better learning I plan to implement random forest, and boosting techniques for probable better solution.

As instructed the task of evaluating models and looking trends for casual and registered rentals separately proved to be a better decision than doing everything over the total rental counts.

The model accuracy scores are still less. To improve them we may try to collect geo spatial data, like rental location, classification of location (like residential area, college are, corporate park), renter's traits, and physically availability factors .

Thanks

Nitansh Gupta