Cab Availability Prediction

ONLINE VEHICLE BOOKING MARKET

Arushi Srivastava | 12/07/2023

Fermi Estimation

Problem statement breakdown:

1. Company background: The team is part of an Online Vehicle Booking Product Startup.

2. Competitive landscape: The startup faces heavy competition from Ola and Uber in the cab booking market in India.

3. Objective: The startup is looking for an alternate segment in the market to gain an early foothold and generate revenue.

4. Task: The team needs to analyze the Vehicle Booking market in India using Segmentation analysis.

5. Strategy development: The team must come up with a feasible strategy to enter the market by targeting specific segments where profitable opportunities exist for offering vehicle booking services.

In summary, the team's goal is to find a new segment in the Indian market where they can establish themselves and generate revenue by offering vehicle booking services. They need to analyze the market using segmentation analysis and develop a strategy to enter the identified segments. This strategy should focus on areas where profitability is possible.

DATA SOURCES:

Provided with an hourly renting data span of two years. Data is randomly divided into **train** and **test** sets. You must predict the total count of cabs booked in each hour covered by the test set, using the information available prior to the booking period. You need to append the **train_label** dataset to train.csv as the '**Total_booking**' column.

datetime - hourly date + timestamp season - spring, summer, autumn, winter
holiday - whether the day is considered a holiday
workingday - whether the day is neither a weekend nor holiday
weather - Clear , Cloudy, Light Rain, Heavy
temp - temperature in Celsius
atemp - "feels like" temperature in Celsius
humidity - relative humidity
windspeed - wind speed
Total_booking - number of total booking

DATA LINKS:
https://github.com/iArushi/FeyNN-Intern/tree/bfbde4a5bf3482815580e60d4e0079c816a396a0/Dataset

DATA PRE-PROCESSING

Libraries used:

- NumPy
- Pandas

STEPS INVOLVED:

- Reading the train dataset and train_label. Merging the train_label with train dataset.
- Checking the data types of each column
- Knowing the unique values in object data types.
- Handling Missing values: No null values in the dataset.
- Converting 'datetime' datatype from string to datetime for extracting hours, month, weekday information.
- Making individual columns for hours, month, weekday from datetime column

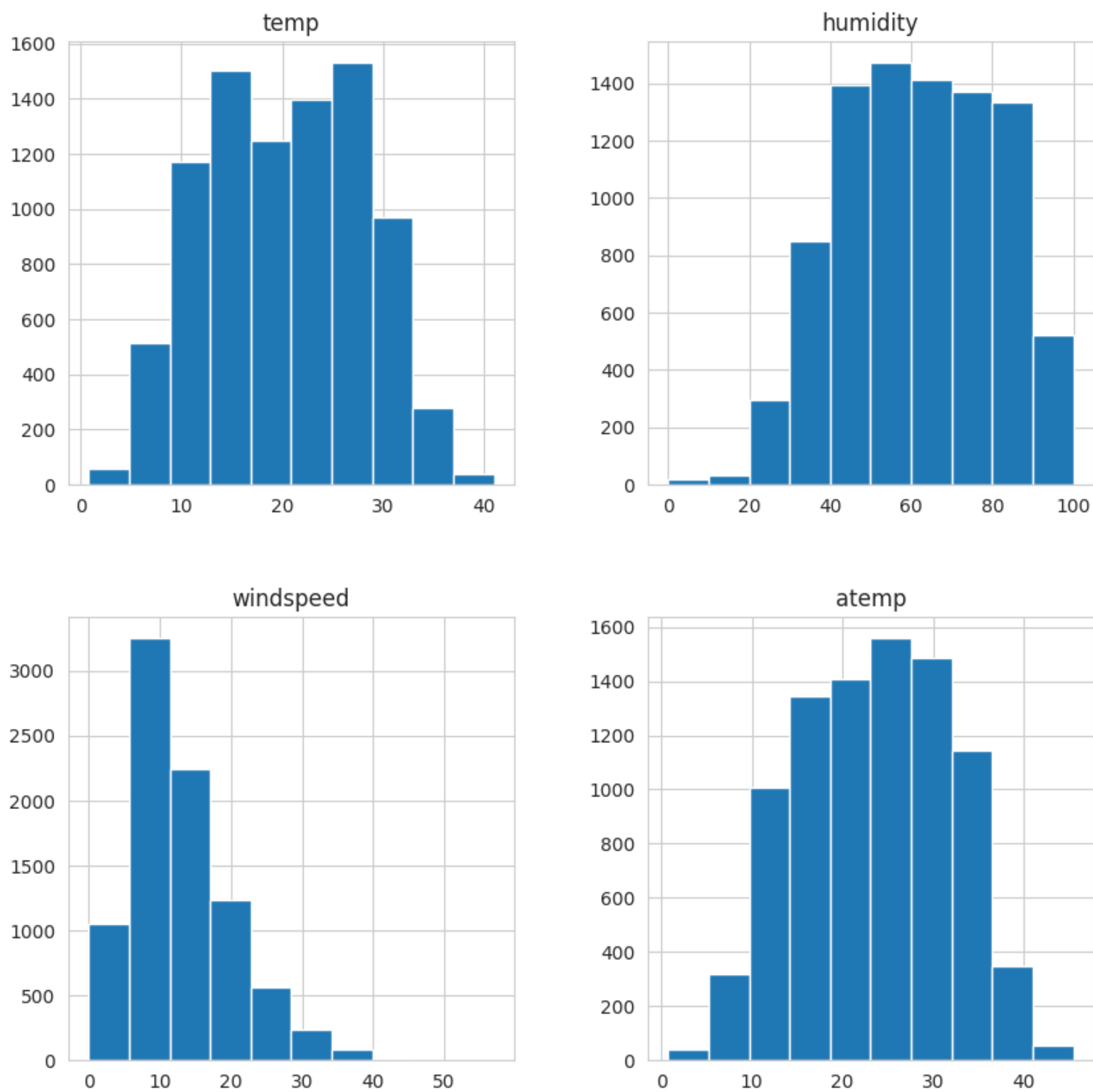
datetim	season	holiday	workingday	weather	temp	atemp	humidity	windspeed	Total_booking	month	hour	weekday
2012-05-02 19:00:00	Summer	0	1	Clear + Few clouds	22.14	25.760	77	16.9979	504	5	19	Wednesday
2012-09-05 04:00:00	Fall	0	1	Clear + Few clouds	28.70	33.335	79	19.0012	5	9	4	Wednesday
2011-01-13 09:00:00	Spring	0	1	Clear + Few clouds	5.74	6.060	50	22.0028	139	1	9	Thursday
2011-11-18 16:00:00	Winter	0	1	Clear + Few clouds	13.94	16.665	29	8.9981	209	11	16	Friday
2011-09-13 13:00:00	Fall	0	1	Clear + Few clouds	30.34	33.335	51	19.0012	184	9	13	Tuesday

SEGMENT EXTRACTION

Following tasks are performed for building the efficient model and executing the project-

TASK 1-

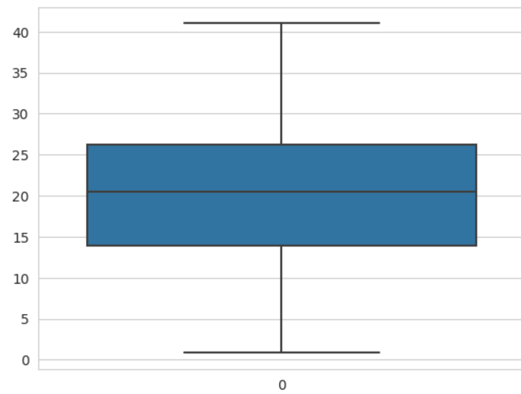
1. Visualize data using different visualization to generate interesting insights.



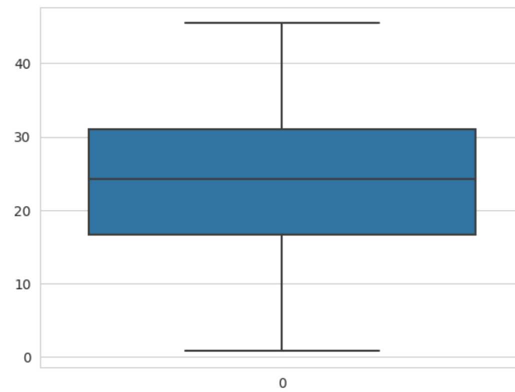
2. Outlier Analysis

OUTLIER>

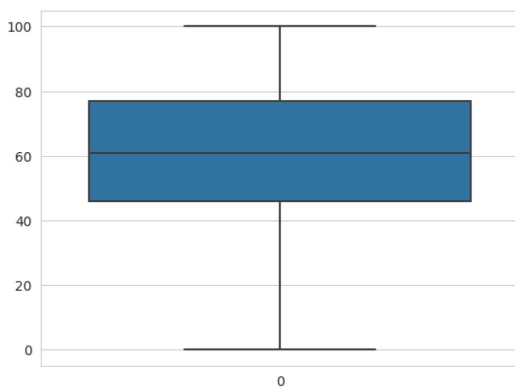
Temp>>



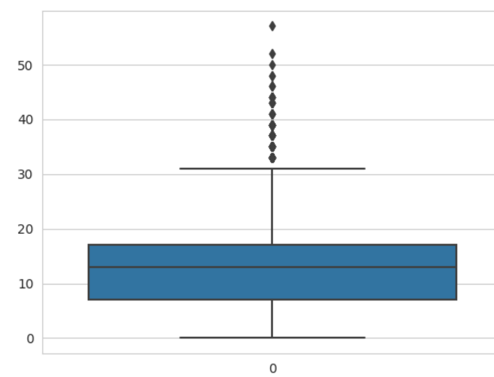
atemp>>



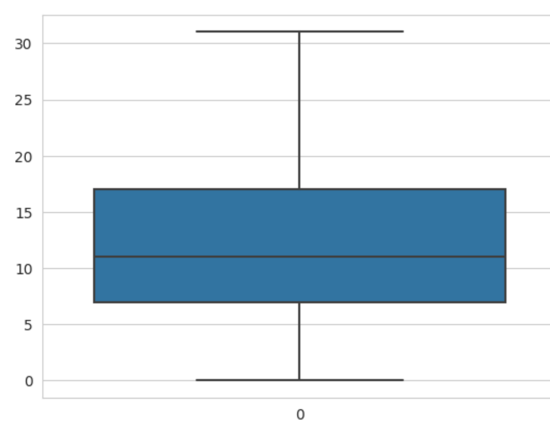
Humidity>>



windspeed>>



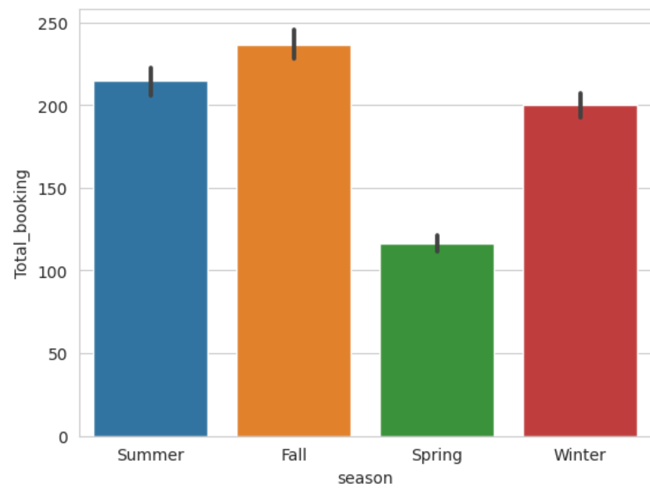
After removing outliers of windspeed using IQR>



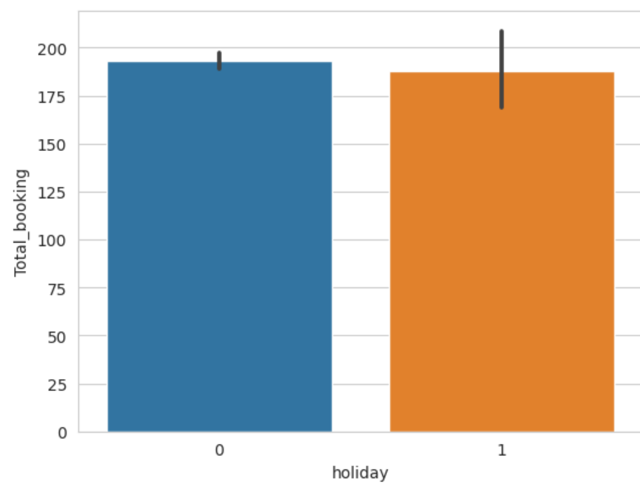
3. Missing value analysis
4. Visualizing Total_booking vs other features to generate insights.

(i) Bar plot

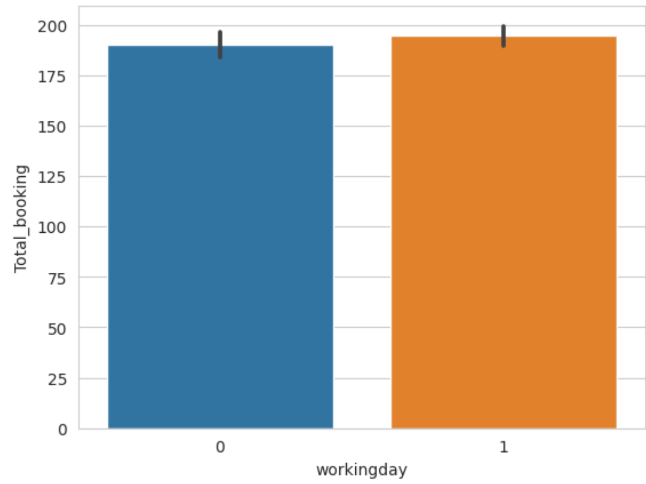
`x='season',y='Total_booking'`



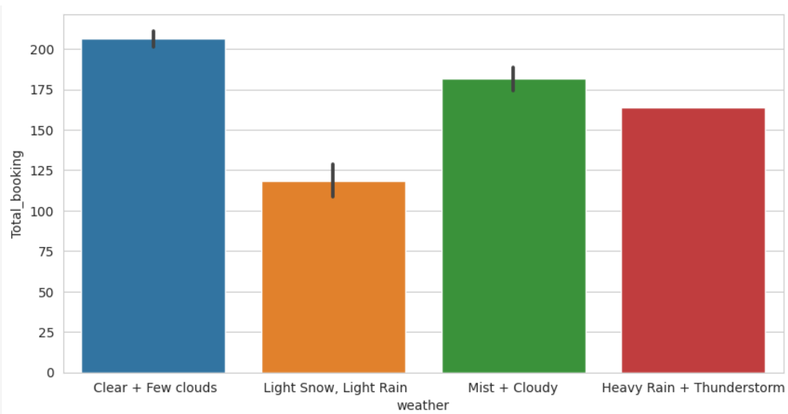
`x='holiday',y='Total_bookin'`



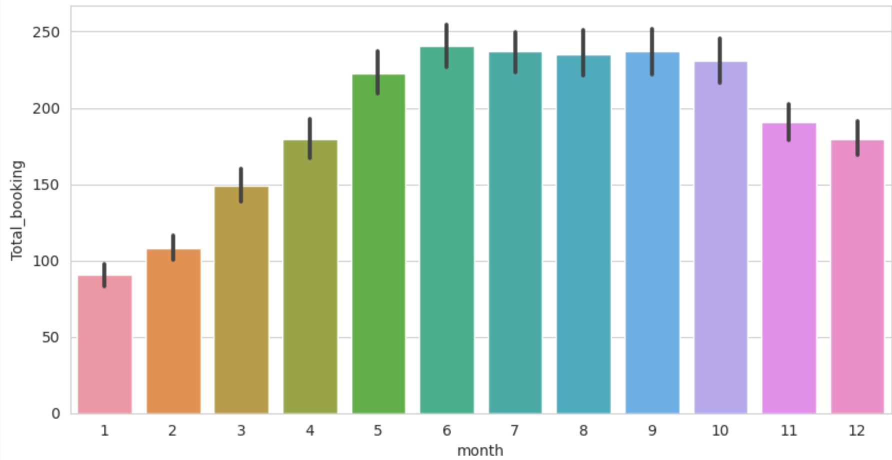
`x='workingday',y='Total_booking'`



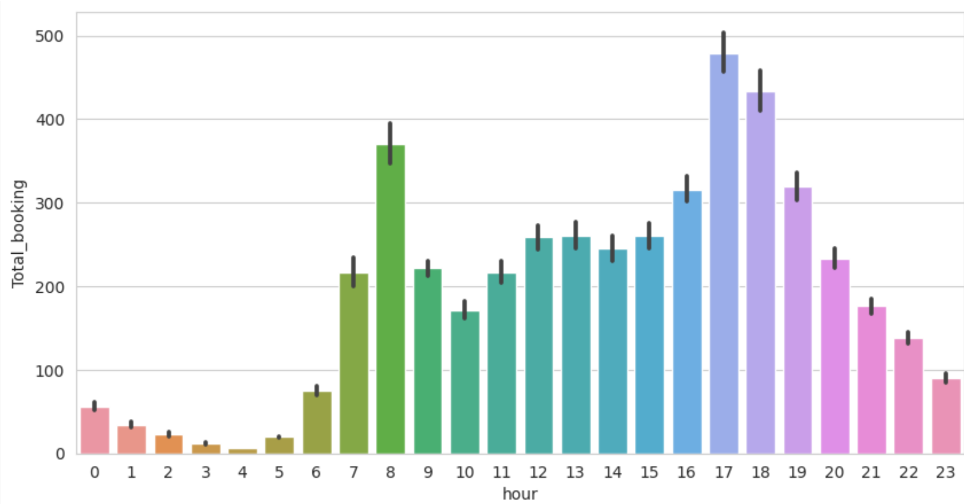
`x='weather',y='Total_booking'`



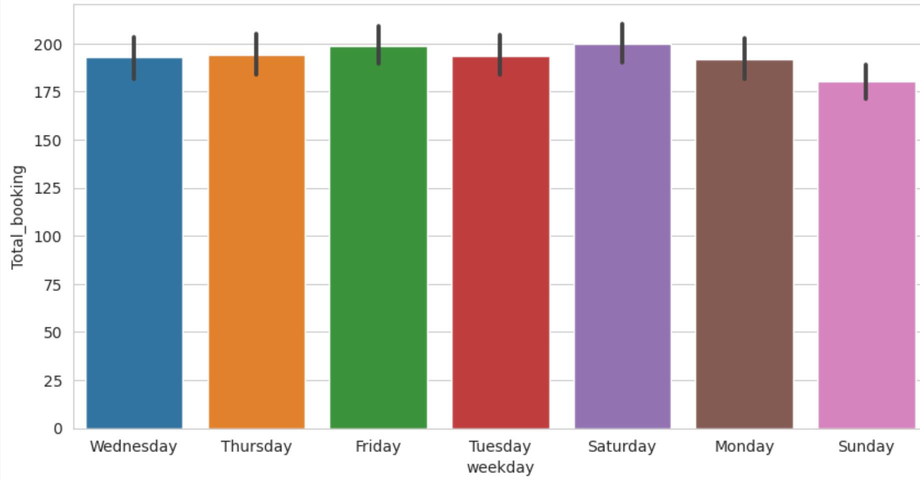
x='month',y='Total_bookin



x='hour',y='Total_booking'

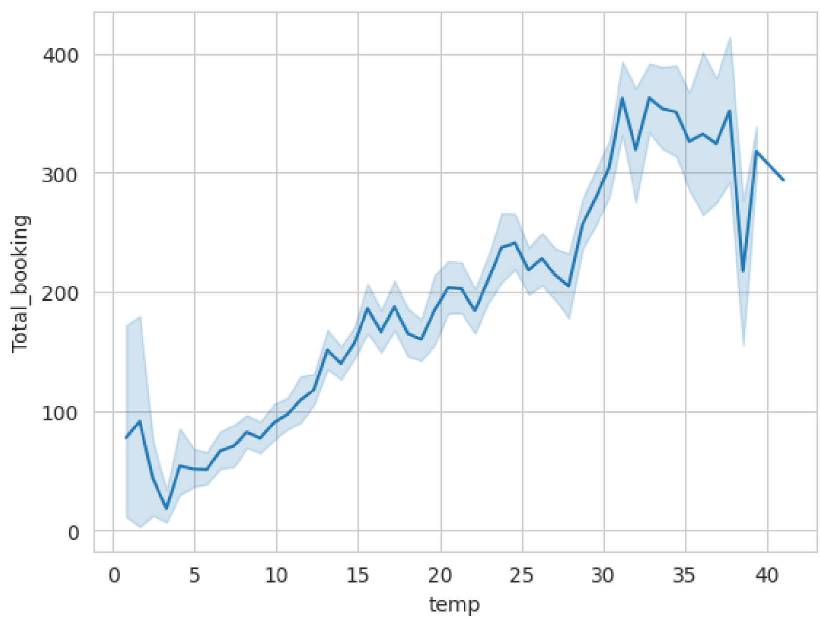


```
x='weekday',y='Total_booking
```

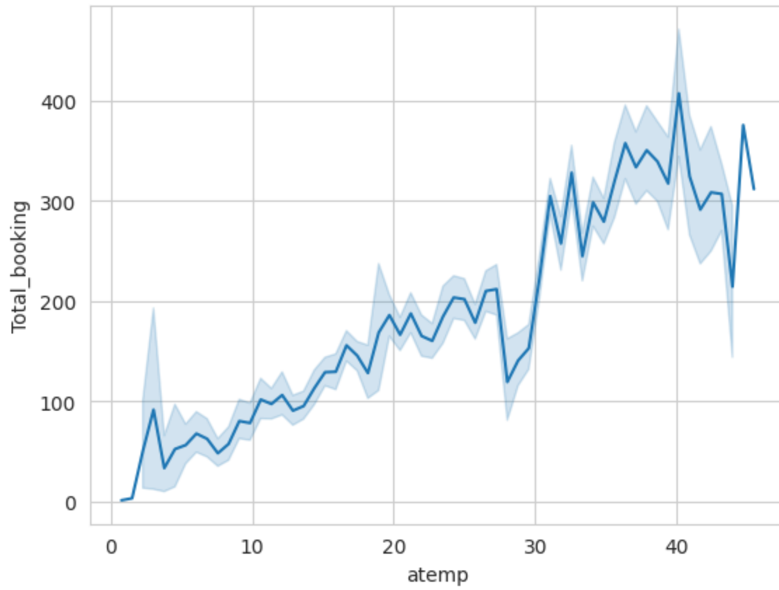


(ii) Line plot

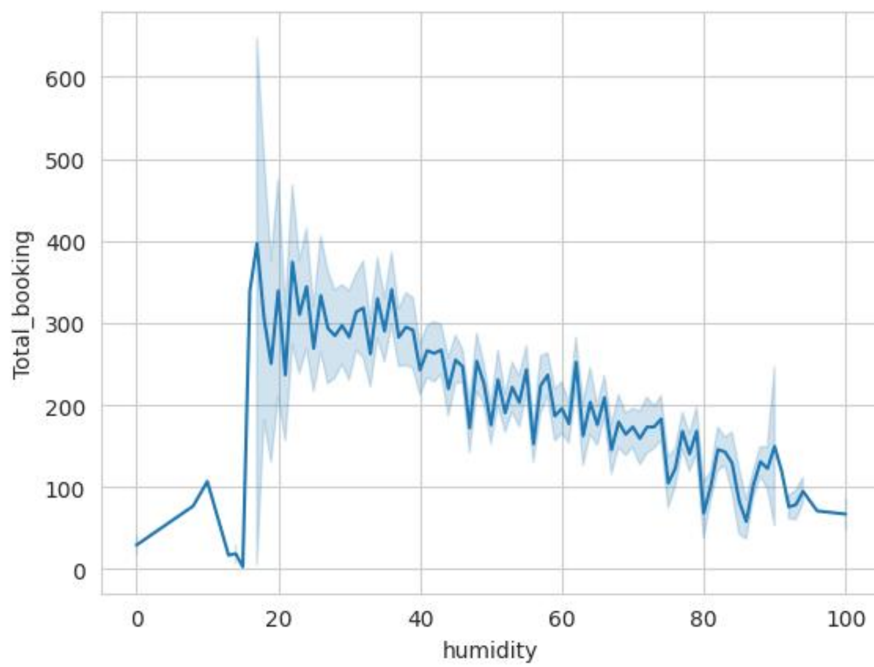
```
x='temp',y='Total_booking'
```



```
x='temp',y='Total_booking'
```

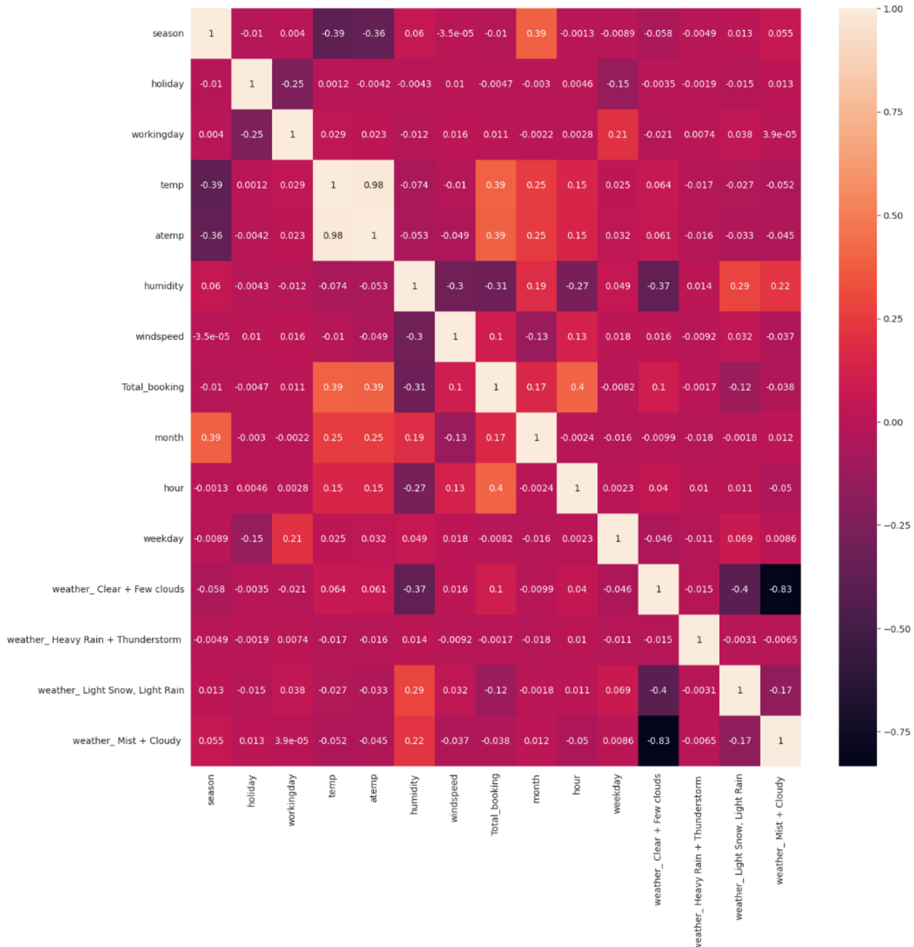



```
x='humidity',y='Total_booking'
```



5. Correlation Analysis

`Heatmap>>`



TASK 2-

1. **Feature engineering**
2. **Grid Search**

   GridsearchCV --> for finding the best parameters through hyper-parameter tuning.

   a) Random forest

   RandomForest-RMSE value:66.519,R2 score:0.859

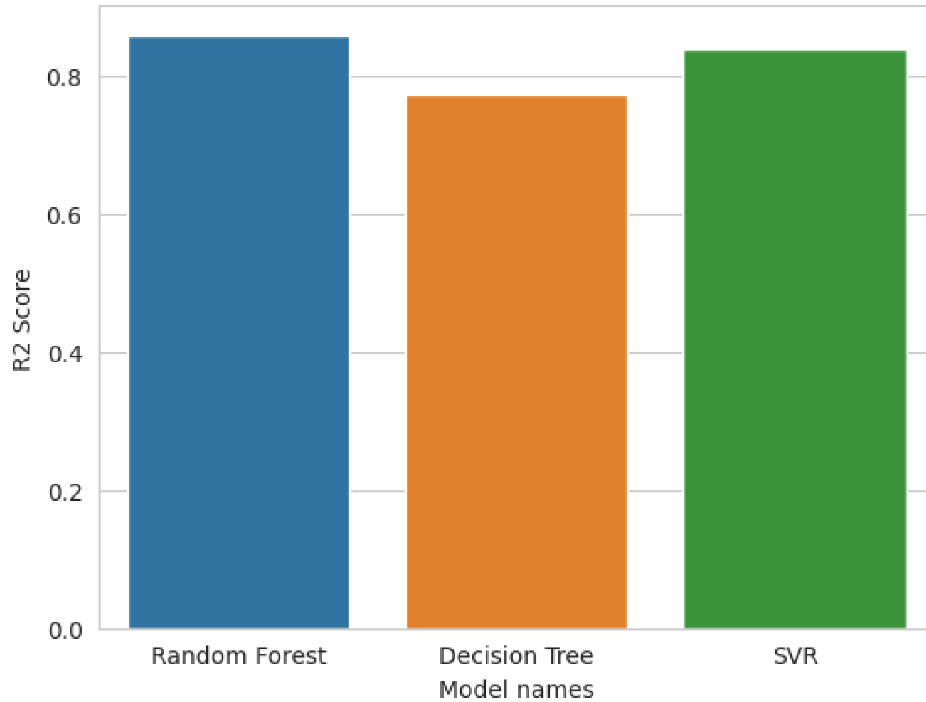
   b) Decision Tree

   DecisionTree-RMSE value:84.15,R2 score:0.774

   c) SVR

   SVR-RMSE value:71.279,R2 score:0.838

d) Plot of different models' performance



Inference :
Random forest model performs well after applying grid search CV also

3. **Regression Analysis**
4. **Ensemble Model**

LinearRegression-RMSE value:139.86,R2 score:0.375
Randomforest-RMSE value:65.066,R2 score:0.865
DecisionTree-RMSE value:65.124,R2 score:0.864
SVR-RMSE value:140.791,R2 score:0.367

Inference:

In bagging –> Random forest as base estimator is performing well compared to other estimators

Inference:

In Adaboost –> Random
Forest as base estimator is performing well compared to other estimators

Final Inference (based on model estimation on 'train.csv' data):

1) Based on regression analysis(without hyper-parameter tuning) --> Random Forest model performs well compared to others

2) Grid search CV (with hyper-parameter tuning) --> Random Forest model performs well compared to others

3) Ensemble learnings --> Both decision tree & random forest models as base-estimators are performing well

Based on above all models 'bagging with random forest as base estimator' is giving best results compared to others on 'train.csv'.

Finally, one of the best models has been chosen for predicting the total cab booking of test data.

Having been checked the model performance by dropping the 'working day, holiday, weekday' features and without dropping the 'working day, holiday, weekday' features.

Based on this, without dropping the 'working day, holiday, weekday' features is giving best R2-Score compared to dropping of 'working day, holiday, weekday' features.

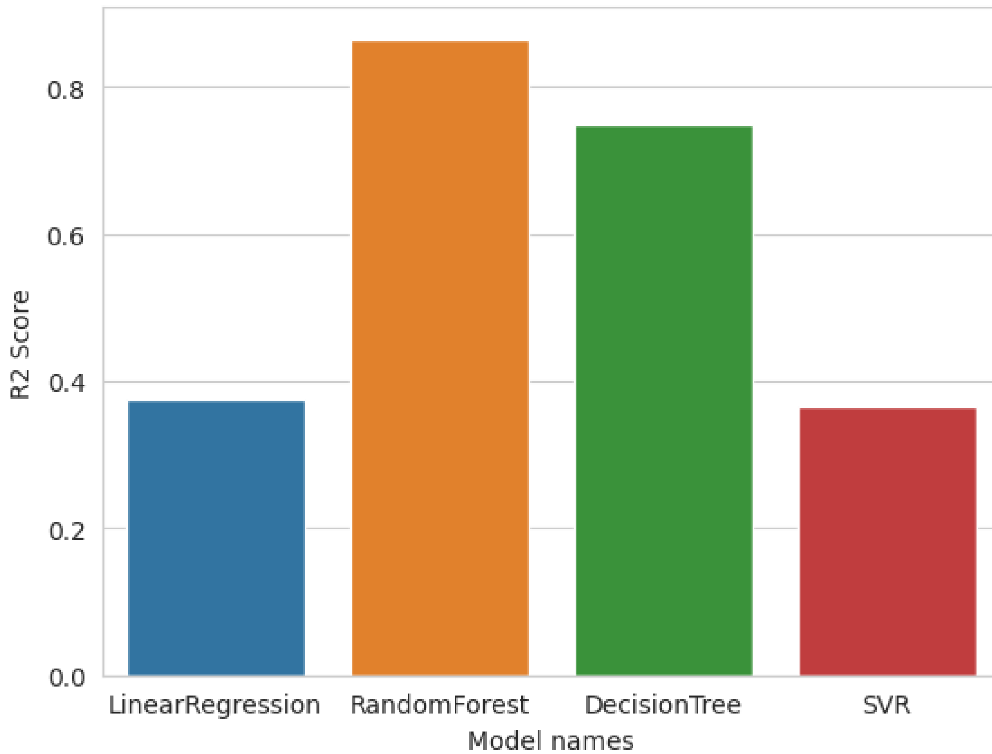
So, for 'test.csv' prediction here i am using without dropping the 'working day, holiday, weekday' features model.

we are going to use 'bagging with random forest as base estimator' for predicting the 'Total_booking' in 'test.csv'.

MODEL BUILDING

For regression analysis, following models have been considered:

1) Linear regression
2) Random Forest
3) Decision Tree
4) SVR



Inference :

Random forest model performs well compared to the remaining models

TARGET

The target segment of cab booking prediction ML algorithms encompasses the potential customers who are seeking cab services. This segment typically includes individuals or groups who require transportation for various purposes, such as commuting, airport transfers, sightseeing, or any other travel-related needs. ML algorithms can be applied to predict and cater to the preferences and requirements of these customers efficiently.

Within the target segment, specific factors can further refine the focus of ML algorithms. Some potential factors to consider when targeting cab booking prediction algorithms include:

1. Demographic Factors: ML algorithms can be designed to consider demographic information such as age, gender, income level, and occupation. Different demographic groups may have distinct travel patterns and preferences, and ML algorithms can utilize this information to tailor recommendations and predictions accordingly.

2. Location-Based Factors: ML algorithms can take into account the geographic location of the customers, including their pickup and drop-off locations. By considering location-based factors such as traffic conditions, distance, and availability of cabs in specific areas, algorithms can provide accurate predictions and recommendations to customers.

3. Timing and Scheduling: Timing is a crucial aspect of cab booking, and ML algorithms can consider factors such as date, time of day, and day of the week. By analyzing historical data and patterns, algorithms can predict peak hours, anticipate high demand periods, and suggest optimal pickup times for customers.

4. Preferences and Special Requirements: ML algorithms can incorporate customer preferences and special requirements, such as vehicle type preferences (e.g., sedan, SUV, or luxury cars), accessibility features, pet-friendly services, or child safety seats. By considering these preferences, algorithms can provide personalized recommendations and increase customer satisfaction.

5. Weather Conditions: Weather can significantly impact travel patterns and demand for cab services. ML algorithms can integrate weather data to predict how weather conditions (e.g., rain, snow, extreme temperatures) may affect the availability and demand for cabs. This enables algorithms to make more accurate predictions and adapt to changing weather conditions.

It's important to note that the specific target segment and factors considered may vary based on the business goals, available data, and the scope of the cab booking prediction system. ML algorithms can be trained and tailored to meet the requirements of the targeted segment, resulting in improved customer experience, optimized operations, and increased efficiency in the cab booking process.

CUSTOMIZATION AND BRIEFING THE MARKET MIX

Customizing and briefing the market mix involves tailoring the elements of the marketing mix to meet the specific needs and preferences of the target market. The marketing mix consists of four key components: product, price, promotion, and place (distribution). Here's how you can customize and brief each element:

1. Product: Analyze the target market to understand their requirements and preferences. Customize the product or service offering to align with their needs. Consider factors such as features, design, quality, and packaging. Ensure that the product satisfies the target market's demands and differentiates itself from competitors. Brief the product team on the desired specifications and features to meet the target market's expectations.

2. Price: Determine the pricing strategy based on the target market's willingness to pay, affordability, and perceived value of the product or service. Consider factors such as competitors' pricing, market demand, and price sensitivity of the target market. Brief the pricing team on the pricing objectives, target price ranges, and any specific pricing tactics or discounts to be applied.

3. Promotion: Develop a promotional strategy that effectively reaches and engages the target market. Customize the messaging, channels, and tactics based on the target market's demographics, preferences, and media consumption habits. Brief the marketing and advertising teams on the key messages, media channels, creative elements, and promotional activities to be executed to effectively communicate with the target market.

4. Place (Distribution): Ensure that the product is accessible to the target market through appropriate distribution channels. Evaluate the target market's shopping behaviors, preferences, and convenience factors. Brief the distribution team on the desired distribution channels, locations, and any specific requirements for reaching the target market efficiently.

In addition to customizing the marketing mix, it's important to continually monitor and evaluate the market response and adjust the strategies as needed. Regularly gather feedback from the target market to ensure that the customized marketing mix is effectively meeting their needs and expectations. Flexibility and adaptability are key in refining the marketing mix to achieve desired results in reaching and engaging the target market effectively.

CODE :

https://github.com/iArushi/FeyNN-Intern/blob/1d52f5fa71b5e988e98fb5d2a221b4c5799284f2/cab_booking.ipynb