Springboard Data Science Course
Capstone Project 2

Time Series Forecasting of Number of Bookings for Camping Sites

By: Priyanka Panhalkar

March , 2025

Introduction

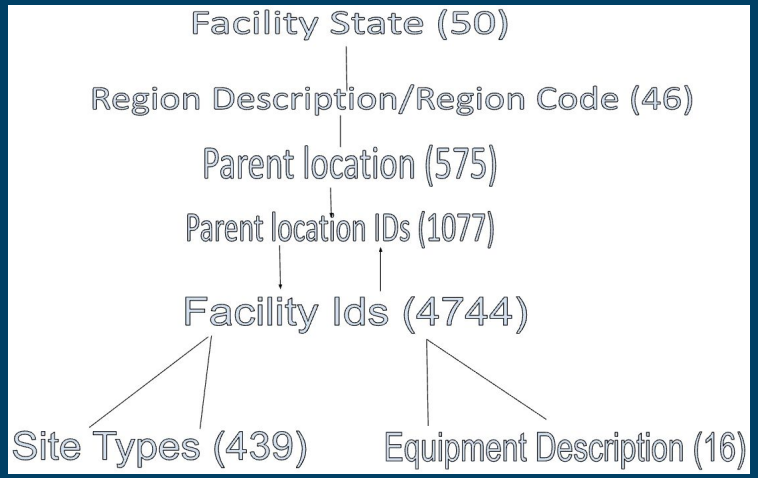
Business Problem: During peak summer season, camping locations would be better prepared to satisfy their customers if they could more accurately predict the demand in advance.

Approach: Using Time Series Forecasting Methods to predict occupation in the Future based on historical information.

Data Collection

Dataset Source: Dataset is available from Link.

Data Wrangling: Dataset was wrangled to find the relationships of columns to each other and a hierarchical map was generated. Eg. total parent locations under Regions.



Data Exploration and Preprocessing

A subset of dataset specific to California was explored and frequency distributions of columns which are relevant to our project goals were performed. Eg the frequency distribution of 'Number of Nights for Camping' (Fig 1) indicates that the majority of bookings were made for 1, 2, and 3 nights.

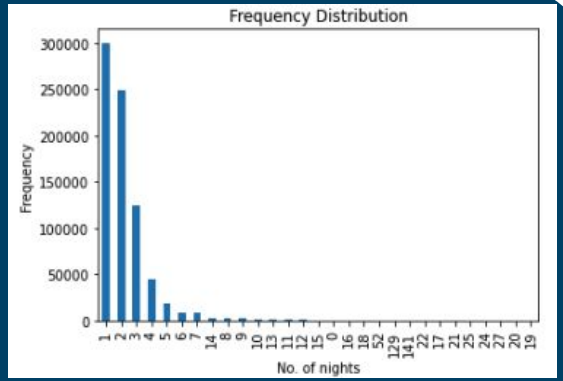


Fig 1

Additional calculated columns were created and plotted.eg delta_order_to_start. It revealed that most of the bookings were done in advance -180 days (Fig2).

Also, original dataset was transformed to individual daily records where each row represents individual row the booking is active. A time series graph for a single facility ID was generated using this dataset. (Fig 3).

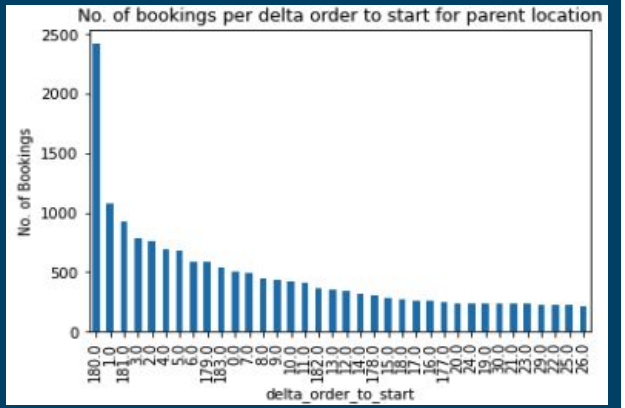


Fig 2

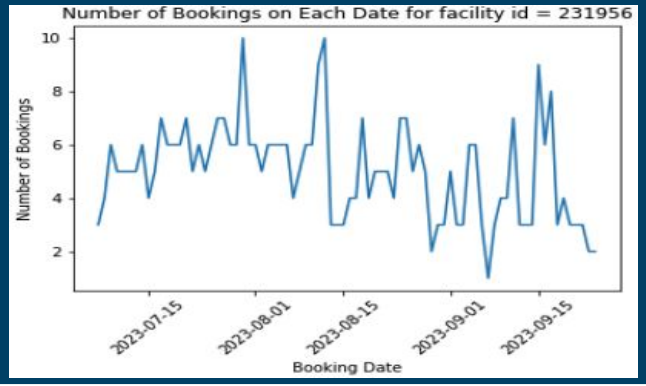


Fig 3

Analysis and Modeling

For Modeling purposes we generated lag features using min,max,mean and stddev of the target column - Number of Bookings on a given date; therefore we transformed a time explicit problem into a time implicit problem that could be addressed using traditional Machine Learning Regression Models.

Time-Preserving splitting of the dataset was performed and three models were evaluated according to MAPE (Mean Absolute Percent Error) on the test set.

The Models we built were based on the following algorithms Linear Regression, Random Forest, and XGBoost.

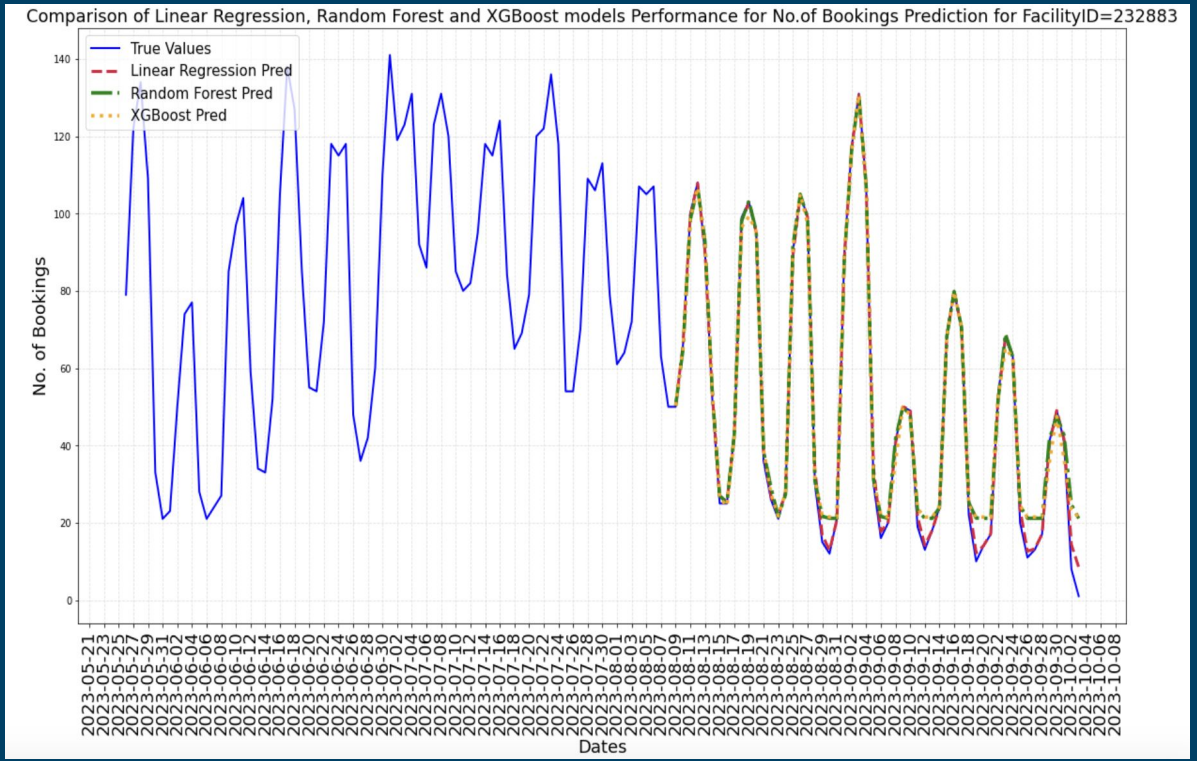
Model	FacID 1	FacID 2	FacID 3	FacID 4	FacID 5
Linear Regression	0.41%	0.97%	2.61%	0.88%	1.8%
Random Forest	13.35%	17.26%	6.68%	2.16%	2.5%
XGBoost	13.42%	17.89%	7.36%	2.71%	3.49%

In final deliverable, we refactored the previous code into reusable functions to generate time series charts displaying actual bookings alongside predictions from different models.

Results and Visualization

Time Series graphs for each Facility ID were generated to observe the behaviour of our models, the one on the right is an example of such graph.

The Linear Regression Model provides overall the most accurate forecasts for Campsite availability, with an average MAPE of 1.33 .



Conclusions and Future Work

- The project aimed at forecasting campsite booking availability to aid in capacity planning, resource allocation, and customer experience optimization for National Forest campsites in California.
- In the Future, we can extend the analysis to other states beyond California to validate the model Performance since our current python codebase is modular.
- A pending enhancement is to introduce hyperparameter optimization by fine tuning Random Forest and XGBoost Models using Grid Search or Bayesian Optimization to consider the possibility of further improve their performance.