# California National Parks Camping Predictions

Jasmine Maquindang, Nima Shafie
*California State University, Northridge*

## I. Introduction

### A. Problem

Camping is a popular outdoor recreational activity for many Americans that generates $166 billion in tourism annually [1, pp.3]. According to the *2021 North American Camping Report*, 48.2 million U.S. households went camping at least one time in 2020, 6 million more than in 2019. Furthermore, 10.2 million (21%) of those households camped for the first time, 5 times the rate of new campers in 2019 [2]. Some of the most sought after destinations for campers are within America's national parks, large areas of land that are federally protected by the National Parks Service (NPS) to preserve its natural beauty, plants, animals, and resources.

Due to growing trends for camping in recent years, it is more important than ever for campers to plan ahead. Most national park campgrounds require reservations through Recreation.gov. Popular destinations, such as Yosemite National Park or Joshua Tree National Park, are often booked weeks or months in advance. Our motive is to aid those planning to camp in national parks within California. Therefore, our project aims to analyze historical reservation data and predict how far in advance a campground was booked.

### B. Data

Datasets for historical online campground reservations were obtained through the Recreation Information Database (RIDB). The RIDB oversees operations for Recreation.gov and provides public data not only limited to national parks, but also information pertaining to historic sites, museums, and other recreational activities throughout the United States. These datasets include booking information such as the campground name, parent national park where the campground is located, date the booking was ordered, date the reservation starts, number of people per campsite, and type of equipment that was used, among other things.

For the purpose of this project, we selected four datasets that span years 2018-2022. The years previous to 2018 are not verified by the RIDB as accurately reported data [3]. Table 1 summarizes the data collected in its raw form. The scope of our study was limited to National Parks managed by the National Parks Service in California. Furthermore, we examined one campground to represent each national park. The largest campground within each National Park was selected, which was determined by the greatest total number of campsites available. Table 2 lists the national parks, campgrounds, and number of available campsites used in this study.

Table 1. Summary of Raw Data

| **Year** | **# Objects (rows)** | **# Attributes (columns)** |
|---|---|---|
| 2018 | 2,712,818 | 57 |
| 2019 | 3,479, 643 | 36 |
| 2020 | 5,114,789 | 37 |
| 2021 | 8,354,633 | 35 |
| **Total** | 19,661,883 | 202 |
| **Average** | 4,915,471 | 40 |

Table 2. Summary of Selected Campgrounds

| National Park | Campground | # Sites* |
|---|---|---|
| Channel Islands | Santa Cruz Scorpion | 31 |
| Death Valley | Furnace Creek | 136 |
| Joshua Tree | Jumbo Rocks | 124 |
| Lassen Volcanic | Manzanita Lake | 179 |
| Pinnacles | Pinnacles | 134 |
| Sequoia and Kings | Sunset | 158 |
| Yosemite | Upper Pines | 235 |

* Number of campsites available according to NPS.gov [4].

### C. Data Preparation

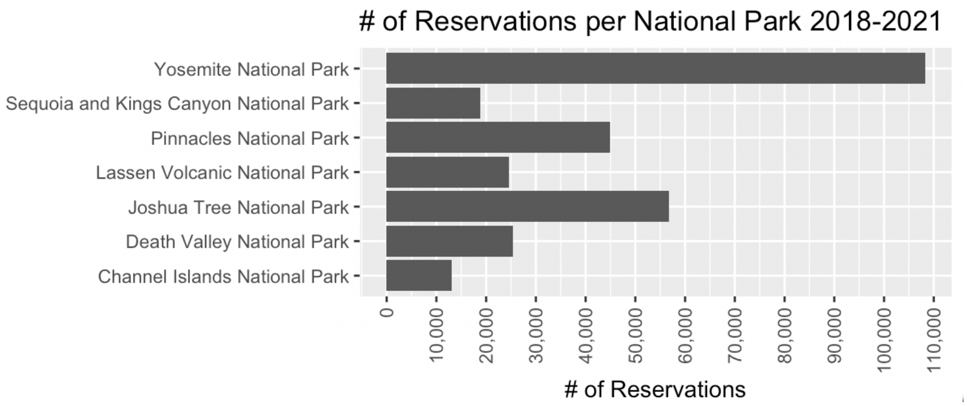
We greatly reduced the number of objects by removing irrelevant data that did not fit into the project scope as mentioned above. Reservations for non-camping activities, reservations for campgrounds outside California, and campgrounds that are not managed by the NPS were deleted. We also removed attributes that do not contribute valuable input to our algorithms, as shown in Table 3. Lastly, it is important to note that reservations can be made up to six months in advance [4]. The raw datasets contained objects with booking horizons greater than 180 days (more than six months) and objects with order dates recorded after their relative start dates; such erroneous objects were removed.

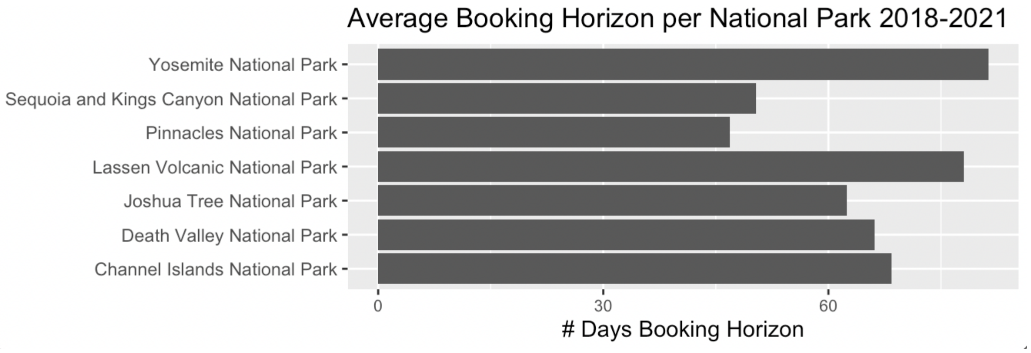
We applied various techniques to handle missing data depending on each attribute. First, we deleted objects if the start date, order date, or end date were empty. We filled in any missing rows for the number of people with that attribute's mean. The dataset for 2018 did not use equipment description as a class. Each equipment type was listed as its own separate attribute, which explains why Table 1 lists 57 attributes above. In order to integrate all the databases together, we merged the attributes for each different type of equipment into one class. If the equipment type was missing or if more than one equipment type was included in a single reservation, we imputed the most frequent value. Once data cleaning was completed, we integrated the four separate years into one finalized dataset. This included 291,884 objects and 8 attributes.

Table 3. Dataset Attributes that were Kept or Removed from the Original Datasets

| Attributes Kept | Attributes Removed |
|---|---|
| parentlocation<br>park<br>sitetype<br>startdate<br>enddate<br>orderdate<br>numberofpeople<br>equipmentdescription | historicalreservationid<br>ordernumber<br>agency<br>orgid<br>codehierarchy<br>regioncode<br>regiondescription<br>parentlocationid<br>legacyfacilityid<br>inventorytype<br>usetype<br>productid<br>facilityid<br>facilityzip<br>facilitylongitude<br>facilitylatitude<br>entitytype<br>entityid<br>facilitytype<br>customerzip<br>customercountry<br>tax<br>usefee<br>tranfee<br>attrfee<br>discount<br>Equipmentlength<br>totalpaid<br>totalbeforetax<br>facilitystate<br>Customerstate |

## D. Data Visualization





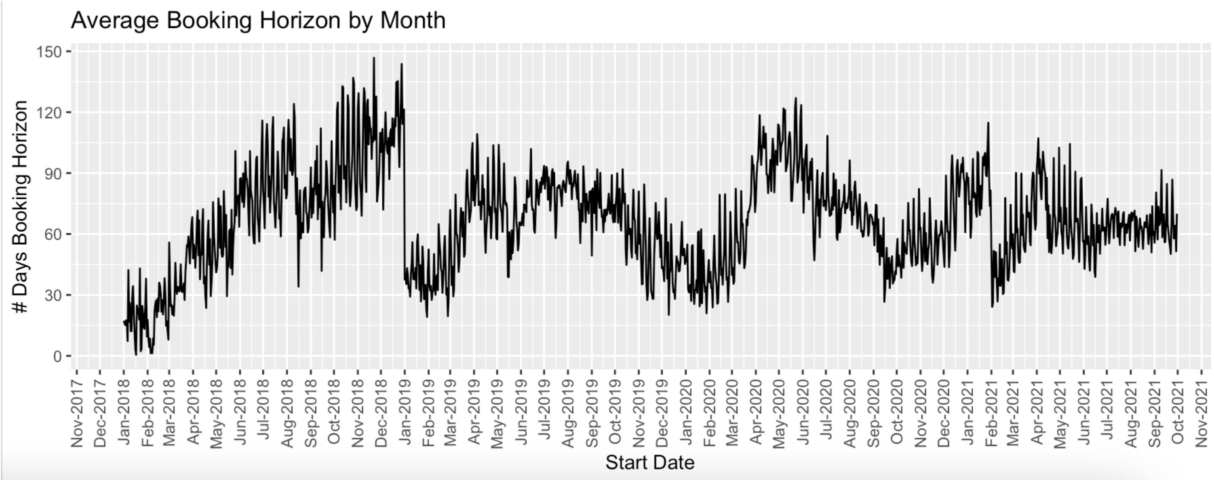
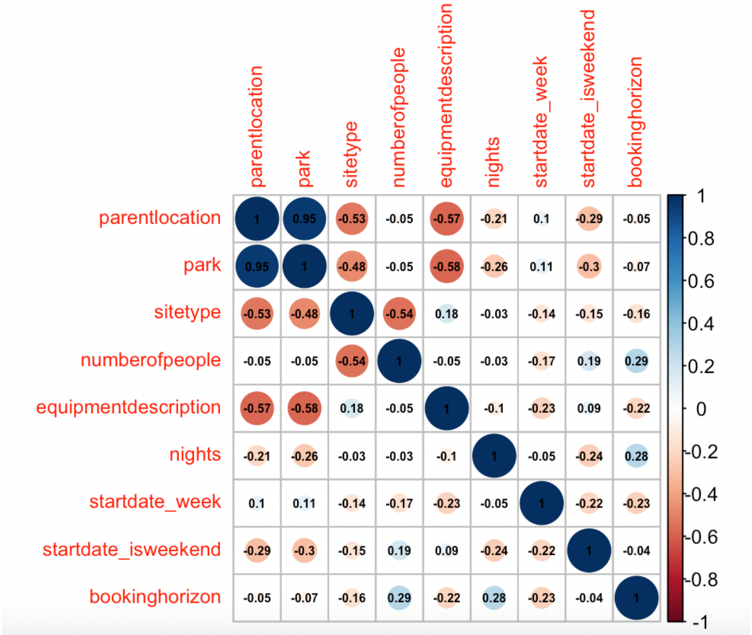


Figure1. Correlation Between Attributes



## II. Related Work

Predictions on reservations for national park campgrounds are understudied, however, we found a similar study that examined forecasting campground demand in national parks. W.L. Rice, et al. [5] recognized that within the topic of tourism demand modeling and forecasting, various methods have been extensively applied to predict the future demand of hotels but not national park campsites. In addition, they claimed that there was no evidence that showed which forecasting method is superior; therefore, they focused on finding the most accurate method suitable for national park campgrounds.

To conduct this study, five approaches that are widely used to forecast tourist arrivals and hotel occupancy were analyzed: moving average (MA), exponential smoothing (ETS), seasonal auto-regressive integrated moving average (SARIMA), neural network autoregression (NNAR), and k-nearest neighbors (KNN), as well as an ensemble method that combined ETS, SARIMA, NNAR, and KNN. Reservation data was acquired from the RIDB for years 2007 to 2017. W.L. Rice, et al. [5] applied these methods to five national parks that were selected based on relatively high percentage number of campsites: Big Meadows Campground in Shenandoah National Park, Elkmont Campground in Great Smoky Mountains National Park, Mather Campground in Grand Canyon National Park, Moraine Park Campground in Rocky Mountain National Park, and Upper Pines Campground in Yosemite National park. They predicted occupancy rates (number of campsites divided by the total number of campsites) at each campground for three, six, and twelve months for each method. To assess the accuracy of their forecasted values, they calculated the mean error, mean absolute error, root mean squared error, and mean absolute percentage error, and compared the forecasted values to the actual values from reservation data.

The result of these models demonstrated that there is no method in particular that is the best for all the campgrounds. They determined that MA was the only method not suitable for forecasting, although other models performed the best for at least one campground because each campground had different characteristics. Thus, they suggest that the ensemble model provides the most accuracy.

In another study, Lee [6] aimed to develop a model to predict future booking arrivals for hotels based on advance booking information and improve short-term (four weeks or less) forecasting accuracy for hotel room demand. First, they discussed three key properties needed for modeling hotel booking arrivals: time-varied arrival rate, high variability in final demand, and a positive correlation between arrival counts in a time partition of a booking horizon [pp. 64]. First, the model should follow the common practice that many industries utilize to observe arrival rates by applying a nonhomogeneous Poisson process (NHPP). Second, it is necessary to incorporate high demand variability into the model because of how volatile and seasonal tourism affects hotel occupancy. Lastly, to increase forecasting accuracy, the model must consider the strong correlations when customers book early in advance or close to arrival date. They developed three Poisson models to capture these characteristics of booking arrivals using daily booking data from 69 major chain hotels across the United States. They examined a short-term booking window of 0-28 days out for 364 arrival dates for each hotel. Rather than testing each arrival date individually, they took 52 samples for each day of the week to determine the best forecasting model.

The first model was a standard NHPP model to estimate an arrival rate function. They observed that booking rates increase on the weekend, in the last week and on the intended stay date, so they incorporated those effects into their function. The second model is a negative binomial model which suggests that within a population of potential customers, one customer  may reserve a hotel room with a success probability at a particular time. They repeat the process until a gamma number of potential customers are remaining in the population. The third model is a negative multinomial model similar to the negative binomial model, however, they assume that each potential customer is independent of one another and a customer may not book at all.

To evaluate their models, they compared the performance of their proposed models with two benchmarks measured for accuracy. They used a regression model (a multiplicative advance booking model) which estimates a booking curve or percentage of accumulated bookings until the arrival time. For the second benchmark, they used their third model, the negative multinomial model to construct a booking curve and get the mean daily arrival rates. They assessed the performance of their models by calculating the Mean Absolute Scaled Error (MASE) and Geometric Mean Relative Absolute Error (GMRAE). The results of Lee's study suggests that the best method varies by hotel and day of the week, however, the negative multinomial model offered the best fit statistically especially for short-term forecasting.

## III. Methods

Our study examines three separate algorithms to capture predictions for the date that a campsite reservation was made prior to a known reservation start date.

### A. Linear Regression

Nima used a 80/20 split for training/testing this supervised machine learning method to predict the exact date a campsite was booked. Linear Regression is a model used for predicting real-value quantities such as an integer. The model makes predictions based on computing a weighted sum of feature inputs and a constant called the bias term. This model was chosen to use on our dataset because the trend of starting dates and order dates were shown to have a positive linear relationship. The raw data has been preprocessed in order to fit this model, since regression requires only numerical data, all categorical values have been converted using a one-hot-encoder method. This method seeks to determine an exact date, where the other two methods will determine a range of dates, thus knowing this, it is expected that this model will have a lower accuracy than the others. Testing the data on a single park from 2018-2021 yields roughly the same results as testing the data on all parks, so it is apparent the amount of data points in this test did not contribute significantly for the linear regression model.

### B. K-Nearest Neighbors

KNN is a nonparametric method used for both regression and classification. In either case, KNN works by assuming data points that are near each other belong to the same class. The value of K is defined as the number of nearest neighbors, the most proximate data points calculated by euclidean distance, to the neighbor to be classified. This algorithm uses a supervised learning approach, which means it learns from labeled training data to predict the outcome of unseen testing data. It is also considered a lazy learning algorithm because it stores the training dataset rather than employing the training data in an explicit training phase. For regression, the output of KNN is a predicted numerical mean average for the target label. In contrast, the output of KNN is the predicted categorial class membership for classification.

Jasmine's initial predictive model used KNN regression in R Studio to target a specific order date. The training and testing data was divided into 80% and 20%, respectively, and included 9 labeled attributes. Some data transformation was necessary before training the model. Firstly, categorical data was converted to integers. Secondly, dates were transformed into the index of the day with respect to 365 days in a year, i.e, 01-15-21 is 15 and

12-7-21 is 341. Min-max normalization was applied to the numerical data such that every value transformed into a decimal value between 0 and 1. Then, Jasmine utilized the “caret” package in R Studio to build the model using the train() function. It is important to determine an appropriate value for K; therefore, she tuned the model parameter controls to utilize k-fold cross validation before running the program on the training set.

Jasmine established a second predictive model using KNN classification to test for higher accuracy. This model accepted similar attributes and followed the same procedure as the KNN regression model, however, rather than predicting an exact date, the model targeted a booking horizon bin (range of dates between order date and start date). Jasmine used binning to label 7 possible time-series classes for the target: > 7 days, > 30 days, > 60 days, > 90 days, > 120 days, > 150 days, and > 181 days. K-fold cross validation was used to find the best value for K once again, specifically 10-folds and 3 repeats.

### C. Random Forest Classification

Similar to KNN, random forest (RF) is also a supervised learning algorithm used for both regression and classification. This method builds an ensemble of decision trees, hence the name “forest”. Each decision tree is generated on sample subsets from the dataset, which are referred to as “bags”, by bootstrapping methods. Furthermore, each decision tree has varying outputs depending on the sample. These outputs are ranked, and the highest are selected, or “bagged”, as the final prediction for the classification label.

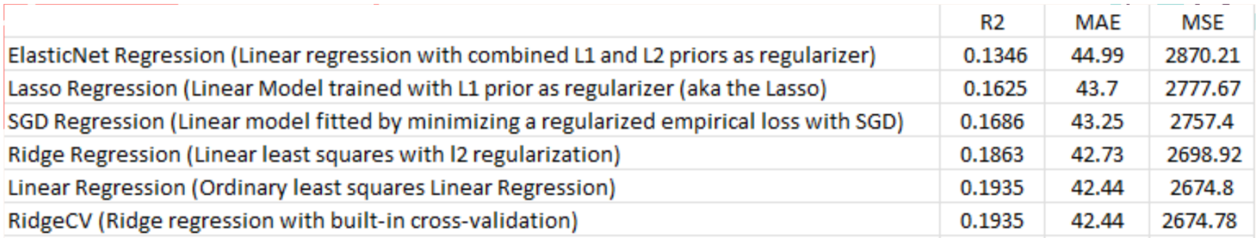
Because KNN is inefficient to use on large datasets, Jasmine established a third model to classify dates that fall within a specific booking horizon bin. She used the “randomforest” package and train() function in R studio to build the model. Most of the procedure was identical to KNN: training and testing data was split into 80% and 20% respectively, the same 9 labeled attributes were utilized, and the same 7 classes of booking horizon date ranges were targeted. Required hyperparameters were selected based on recommendations by Boehmke, B., and Greenwell, B [7]. For the argument “mtry”, which is the number of variables to randomly sample as candidates at each split, they recommend calculating the value by taking the square root of p, where p is the number of attributes fed into the model. In addition, for the argument “ntrees” which is the chosen number of decision trees in the forest, they suggest using p*10. In this case, mtry was 3 and ntrees was 90.

## IV. Evaluation

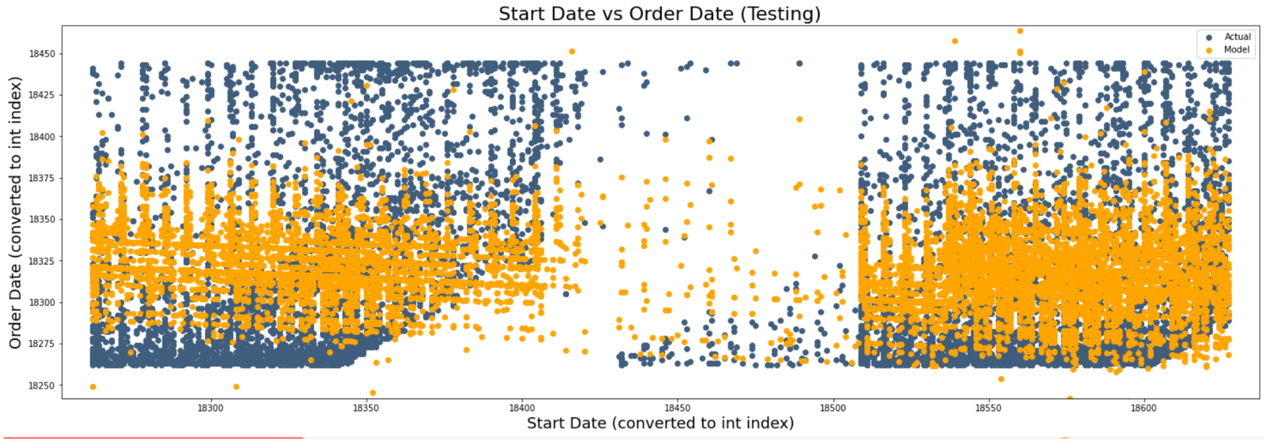
### A. Linear Regression

The evaluation metrics for linear regression are based on a few major variables. The evaluation for determining accuracy is based on the R2 score which is a measure of variance in the predictions explained in the dataset. It is the difference between the predictions made by the regression model and the samples in the dataset. The other two metrics are used to evaluate the errors in the dataset, in this category we used Mean Square Error (MSE) and Mean Absolute Error (MAE). These metrics are considered the loss/cost function that measures the error between the predicted target value (booking_horizon) and the true target value. The metric that was more significantly used to evaluate between the regression model with KNN and RF Classification was the R2 accuracy score.

Figure 2.



| | R2 | MAE | MSE |
|---|---|---|---|
| ElasticNet Regression (Linear regression with combined L1 and L2 priors as regularizer) | 0.1346 | 44.99 | 2870.21 |
| Lasso Regression (Linear Model trained with L1 prior as regularizer (aka the Lasso) | 0.1625 | 43.7 | 2777.67 |
| SGD Regression (Linear model fitted by minimizing a regularized empirical loss with SGD) | 0.1686 | 43.25 | 2757.4 |
| Ridge Regression (Linear least squares with l2 regularization) | 0.1863 | 42.73 | 2698.92 |
| Linear Regression (Ordinary least squares Linear Regression) | 0.1935 | 42.44 | 2674.8 |
| RidgeCV (Ridge regression with built-in cross-validation) | 0.1935 | 42.44 | 2674.78 |



## B. K-Nearest Neighbors

It must be noted that Jasmine was unable to apply the KNN regression model to the entire 2018-2021 dataset with all campgrounds due to high cost in terms of time to process the dataset in R Studio. Instead, a subset containing the Santa Cruz Scorpion campground within Channel Islands National Park was created. The performance metrics for this subset are presented in Figure 3. The output determined that 7 nearest neighbors was the optimal value for K. Since the value of K was small relative to the size of the dataset, the model was overfitting the fed data. This implies that the model worked well with training data but poorly predicted testing data, which is reflected in the RMSE score of 0.22. We assumed similar results would likely be produced if the model was applied to the entire dataset.

The overall performance of the KNN classification score is represented in Figure 4. The R2 accuracy score yielded 44.11%, which indicates that the model successfully predicted nearly half of the testing data. We can infer that noise caused by low correlation between attributes may be a potential contributing factor to underperforming accuracy scores. Furthermore, the confusion matrix fails to show whether the predictions were precise.

Figure 3. Output of k-fold  cross validation on KNN regression model

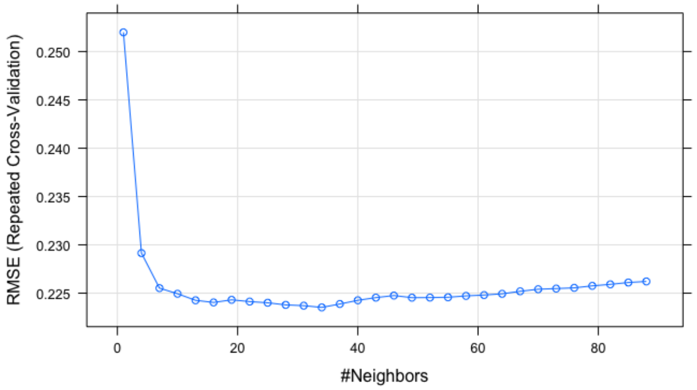
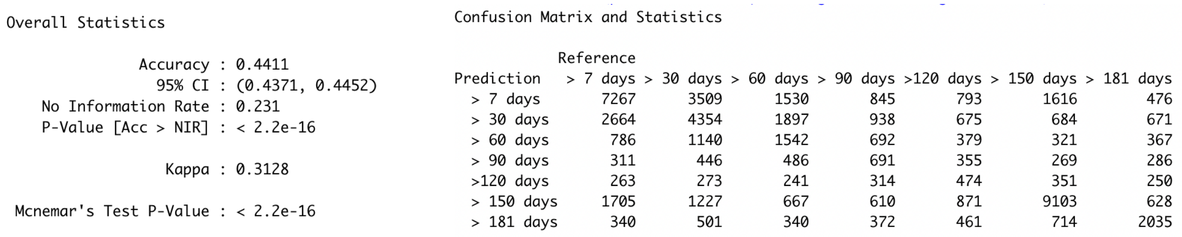


Figure 4. Overall performance of KNN classification model



```
Overall Statistics

               Accuracy : 0.4411
                 95% CI : (0.4371, 0.4452)
    No Information Rate : 0.231
    P-Value [Acc > NIR] : < 2.2e-16

                  Kappa : 0.3128

 Mcnemar's Test P-Value : < 2.2e-16
```

```
Confusion Matrix and Statistics

            Reference
Prediction   > 7 days > 30 days > 60 days > 90 days >120 days > 150 days > 181 days
  > 7 days       7267      3509      1530       845       793       1616        476
  > 30 days      2664      4354      1897       938       675        684        671
  > 60 days       786      1140      1542       692       379        321        367
  > 90 days       311       446       486       691       355        269        286
  >120 days       263       273       241       314       474        351        250
  > 150 days     1705      1227       667       610       871       9103        628
  > 181 days      340       501       340       372       461        714       2035
```

### C. Random Forest

In comparison to the other models, RF was the best performing prediction model in terms of accuracy. Figure 5 demonstrates the error rate for each targeted class over the number of trees in the RF model. The solid black line represents the overall Out of Bag (OOB) estimate of error rate, which is the model's mean error using predictions from trees that do not contain that data point in its respective sample. Each of the other lines, differentiated by color and dashed lines, represent individual OOB error rates for each target class label. The estimated OOB error rate scored 54.33% as shown in Figure 7. Hence, the estimated accuracy rate is 45.67%, slightly higher than the KNN classification model. In addition, it is evident that the error rate for all the classes reaches a plateau around 25 trees. At this point, additional trees have little effect on the prediction accuracy.

Another built-in feature that the RF model tests for is variable importance. In Figure 6 the mean decrease accuracy calculation determines how much accuracy a model loses by excluding each attribute. The most impactful attributes include nights, numberofpeople, startdate_isweekend, and startdate_week. The resulting attributes of this indication are analogous with the correlation matrix in Figure 1; the same attributes have the highest correlation. Lastly, the mean decrease gini shows us how much each attribute decreases node impurity. It is worth mentioning that the startdate_week and parent location have the highest affect on how well a decision tree can split. This confirms that the reservation start date and national park have the most influence on predicting how far in advance a campsite was booked.

Figure 5. RF Model OOB error Output

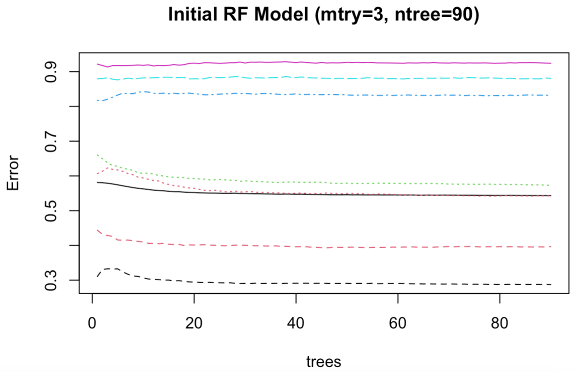


Figure 6. RF Variable Importance Output

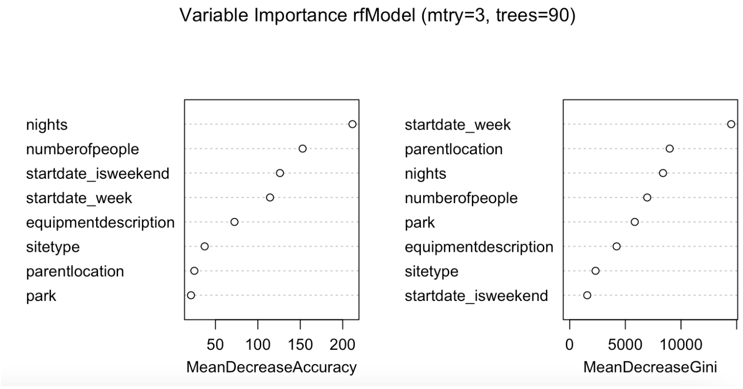
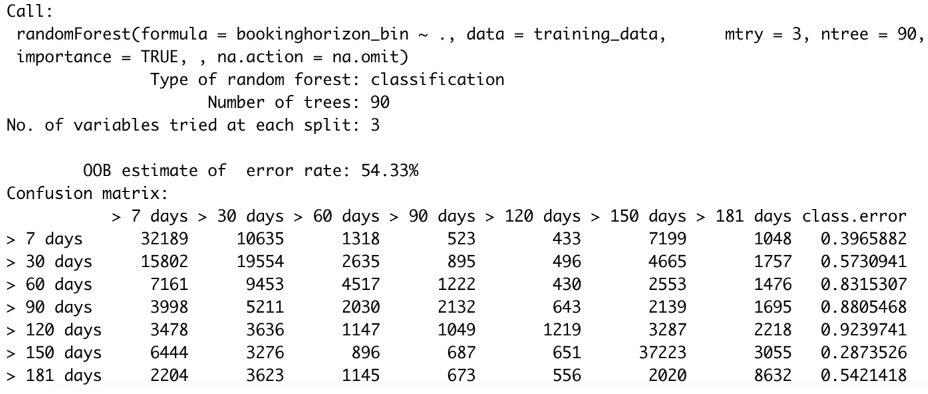


Figure 7. Overall Performance of RF Classification Model



```
Call:
 randomForest(formula = bookinghorizon_bin ~ ., data = training_data,      mtry = 3, ntree = 90,
 importance = TRUE, , na.action = na.omit)
               Type of random forest: classification
                     Number of trees: 90
No. of variables tried at each split: 3

        OOB estimate of  error rate: 54.33%
Confusion matrix:
           > 7 days > 30 days > 60 days > 90 days > 120 days > 150 days > 181 days class.error
> 7 days      32189     10635      1318       523        433       7199       1048   0.3965882
> 30 days     15802     19554      2635       895        496       4665       1757   0.5730941
> 60 days      7161      9453      4517      1222        430       2553       1476   0.8315307
> 90 days      3998      5211      2030      2132        643       2139       1695   0.8805468
> 120 days     3478      3636      1147      1049       1219       3287       2218   0.9239741
> 150 days     6444      3276       896       687        651      37223       3055   0.2873526
> 181 days     2204      3623      1145       673        556       2020       8632   0.5421418
```

## V. Conclusion

The majority of our findings have yielded that the proportion of the largest booking horizon lies within order dates that have been reserved within seven days of the start date, and between four to five months before the start date. We can see the similarity between National Parks and hotels by comparing the data that we mined from the RIDB historical reservation data with the research provided by the related works we have found. In comparison we have used a mixture of supervised machine learning models in an attempt to predict campground reservation order dates.

Given our dataset and the features that we have to work with, the three different models that we implemented had their shortcomings. The challenges lie within the correlation between data that was presented by the RIDB database. The KNN and RF algorithms presented a more accurate prediction but at the cost of lengthy run times and a high computational cost. The Linear Regression model ran within a fraction of the time, however performed the most poorly in terms of accuracy, this model proved to be less than half as accurate as the other two. The KNN and RF models were designed to predict a customizable range of days that the booking horizon can fall in, whereas the linear regression model was purely predicting the exact days of the booking horizon.

The impact from the low amount of correlated features dampened our model accuracy and increased the mean average/square error that was introduced in the predictions that the models made. Even with the addition of the few highest correlating features that were added by conducting feature engineering our models failed to achieve an acceptable accuracy. Ultimately this leads us to conclude that the historical campground reservation data that we worked with is indeed independent from predicting future campground reservations.

## VI. Futurework

The next steps that are to be taken to improve on campground predictions involve factors that are prevalent to campground bookings, some of these factors involve circumstances that are beyond our control. The first major factor that we saw in our data included seasons of the year that promoted a more enjoyable camping experience. Our data shows that there was a higher influx of reservations during summertime, and the shoulder season, more specifically, April through August. Giving different weights and tweaking hyperparameters based on seasonality may yield better results.

Although difficult to predict, natural disasters can majorly impact campground predictions. California is known for wildfires and having a system that can read live updates regarding such natural disasters can change the course of predictions to more closely acclimate to the restriction of campground bookings. The latest example can be taken from the uptrend in camping that was brought on by the COVID pandemic.

Introducing a different style of learning can also potentially strengthen the model prediction. Training the system with an incremental learning technique can help keep the predictions on track. Online learning would take advantage of the large dataset from the RIDB provided there be a continuous flow of campground reservations transferred to the machine learning model. This technique can substantially lower the cost and resources that are needed to run our more computational and time intensive models such as KNN and RF.

## References

[1] National Park Service. *Campground Industry Trends*. US Department of the Interior. 2020. [Online]. Available: https://www.nps.gov/subjects/policy/upload/CBRE_Campground_Trends_Report_3-16-20.pdf

[2] Kampgrounds of America [Online]. Available: https://outdoorrecreation.wi.gov/Documents/Research%20Library%20Page%20files/US%20-%20Demographics%20%26%20Participation/2021-north-american-camping-report.pdf

[3] RIDB/Recreation.gov Historical Reservation Data. Recreation.gov, 2018-2022. [Online]. Available: https://ridb.recreation.gov/download

[4] "Nps.gov (U.S. National Park Service)," National Parks Service. [Online]. Available: https://www.nps.gov/index.htm.

[5]Rice, W. & Park, S., Pan, B, et al. "Forecasting Campground Demand in US National Parks," *Annals of Tourism Research*, Elsevier, 2019, vol. 75(C), pp. 424-438. [Online]. Available: https://www-sciencedirect-com.libproxy.csun.edu/science/article/pii/S0160738319300131

[6] Lee, M. "Modeling and forecasting hotel room demand based on advance booking information"
*Tourism Managemen*t. 2018, pp. 66. 62-7. [Online] Available: 1https://www.sciencedirect.com/science/article/pii/S0261517717302431

[7] Boehmke, B., Greenwell, B. "Random Forests" in *Hands-On Machine Learning with R,* Taylor & Francis Group, 2019. [Online]. Available: https://bradleyboehmke.github.io/HOML/

**Database Link**

**Code**

All the code you wrote with comments that explain your implementation.

| | |
|---|---|
| Nima's Code Link | Jasmine's Code Link |