African Journal of Hospitality, Tourism and Leisure. ISSN: 2223-814X

April 2022, Vol 11, No 2, pp. 457-474

Factors Affecting the Visitation to National Parks Using Machine Learning Techniques: The Case of National Parks in Rwanda

Musonera Abdou*
African Centre of Excellence in Data Science, College of Business and Economics, University of Rwanda, Kigali, Rwanda, Email, abdoumusonera@gmail.com

Edouard Musabanganji
School of Economics, College of Business and Economics University of Rwanda, Kigali, Rwanda, Email, musabanganji@gmail.com

Herman Musahara
School of Economics, College of Business and Economics University of Rwanda, Kigali, Rwanda, Email, hmusahara@gmail.com

Corresponding Author

How to cite this article: Abdou, M., Musabanganji, E. & Musahara, H. (2022). Factors Affecting the Visitation to National Parks Using Machine Learning Techniques: The Case of National Parks in Rwanda. African Journal of Hospitality, Tourism and Leisure, 11(2):457-474. DOI: **https://doi.org/10.46222/ajhtl.19770720.236**

Abstract

The current study set out to identify factors affecting the number of visitors to national parks using machine learning techniques. The results of different linear regression and random forest models on both the train and test sets were compared using RMSE and R^2. Taken together, both random forest and linear regression models were able to predict better on the train set, but all failed to make better predictions on the test set. Both linear regression and random forest models performed better using data from Akagera National Park than Volcanoes and Nyungwe National Parks. The most important features to explain the number of visits to national parks include price, park-specific characteristics, and different months of the year whose features tend to vary from one park to another. This implies that forecasting future visits to different national parks will not only allow policy makers and the park management to make effective planning and efficient allocation of resources, but will also provide valuable information to various people as they plan to visit various national parks.

Keywords: Machine learning; gorillas; Akagera; Nyungwe; Rondom forest

Introduction

National parks and other protected areas in different tourism destinations have common feature to attract overseas visitors and making a measure contribution towards improved livelihoods of the communities and economic development (Thomas & Koontz, 2021; Halpenny, 2010; Clark et al., 2019; Manning, 1980)**.** Direct economic value created include economic impact beyond users, immediate use, saving or price. Indirect value consists of impact on property value, reduce water pollution, reduced air pollution, community cohesion and health. In other words, revenues accruing from national parks benefit two main economic agents that is individuals and the government in the form of income and savings (Harnik & Crompton, 2014)**.** Though National parks are an important source of revenues, its main purpose of preserving the natural landscape and biodiversity should prevail. Therefore, a balance should be maintained between the provision of recreation facilities or revenue generation and the allocation of resources towards the creation and conservation of National Parks (Schägner et al., 2016)

The identification of park characteristics and factors affecting the demand for products and services offered by national parks lay a foundation for the effective prediction of the impact

of management decisions on the total number of visits to national parks as well as the overall social and economic impact that result from change in demand for products and services offered by National Parks (Neuvonen et al., 2010**).** While the emergence of national parks and recreation facilities results in increased business opportunities, there are several economic, ecological factors and social factors to be taken into consideration in order to maintain high quality services offered to visitors. A National Park designation is more likely to cause the increase in visitors, shifting demographics and behavior of the visitor, activity involvement and improved positive attitude towards the national parks to be visited (Fredman et al., 2007)

Previous authors have attempted to identify drivers of national park visitation, length of stay and willingness to pay for services offered by national parks using various approaches. Forecasting techniques that have been increasingly attracting the interest of researchers include Regression techniques (de Castro et al., 2015); (Scholtz et al., 2015); static and dynamic econometric methods (Wang et al., 2021; Rice et al., 2019); AI-based models (Clark et al., 2019; (Rice et al., 2019); Hybrid models ((Rice et al., 2019); Factor analysis (Mutanga et al., 2017; Kim et al., 2003**;** Kruger& Saayman, 2010)**.** However, little attention has been paid to the application of machine learning techniques to predict the effects price and other park characteristics on the visitation to national parks. Therefore, the current study applies machine learning techniques to identify drivers to visit national parks in Rwanda. The remainder of the paper is organized as follows: Section 2 dives into the literature review while section three describes the methods and techniques used. Section four discusses the findings of the study, and the section covers related conclusions.

Literature review

The main mission of protected areas is twofold. They are supposed to play a critical role in the conservation of natural environment and the provision of recreation opportunities (Leung et al., 2018). Based on visitor experience, recreation facilities can be grouped into four categories that are settings, experience, recreation facilities and benefits (Manning, 2011). Settings are within the control of the managers of the protected areas while benefits and experiences result from visits. Settings fall into three main categories: physical, social and managerial elements. Recreation facilities are the outcomes of the combination of settings. Recreation opportunities include trekking, picnics and other facilities at disposal of various groups that visit protected areas/national parks (Mutanga et al., 2017)

Researchers have used different regression techniques to determine factors affecting the visitation to national parks and the length of stay. Regression techniques that were predominantly used include Generalized Methods of Moment **(**Poudyal et al., 2013)**;** Correlation and linear regression (Morgan et al., 2011**;** Smeral, 2009); Panel data regression (Kim & Song, 2017); (Meric & Hunt,1998; Nerg et al., 2012); Spatial regression (Hanink& Stutts, 2002). The main determinants that were identified include park characteristics and quality of products offered (de Castro et al., 2015; Schägner et al., 2016), park pricing policies (Schwartz & Lin, 2006)**,** demographic and behavioral characteristics of the visitors (Kim &Song, 2017**;** Meric & Hunt, 1998) financial and non-financial shocks (Morgan et al., 2011; Smeral, 2009), climate conditions (Schägner et al., 2016), distance and travel costs (Hanink&White, 1999**;** Manfredo et al., 1996). However, researchers appear to have overlooked the comparison of the results of linear models with AI-based models such as a random forest.

More recently, several attempts have also been made to predict visits to national parks using various econometric and machine learning techniques (Clark, 2019; Wang et al., 2021; Rice et al., 2019); Poudyal et al., 2013). For Example, Clark et al. (2019) used google trends to forecast visits to national parks. Though the google trend performed better than Vector

Despite parks differences, Vector Autoregressive model was better than google trend model for some parks. Google trends like other such engines need to be used together with other forecasting methods in order to yield better results. Researchers also used forecasting approaches such as moving average, exponential smoothing, seasonal auto-regressive integrated moving average and neural networks to predict camping demand (Rice et al., 2019). However, owing to the unique features of the campsite demand such as seasonal availability of campsites, lack of substitutes, the of use modern forecasting techniques such k-nearest neighbours and ensemble methods becomes inevitable in order to increase accuracy in predicting campsite demand.

In recent years, there has been an increasing amount of literature on various factor analysis approaches to determine pull and push factors to visit national parks and other nature-protected areas (Reihanian et al., 2015; Smeral, 2009); Zydroń et al., 2021; Mutanga et al., 2017; Kruger & Saayman, 2014). Mutanga et al. (2017) applied principal component analysis (PCA) to investigate Tourists' travel motives for visiting national parks, tourists' wildlife tourism experiences and predictors of wildlife experience, and visitors' overall satisfaction with trip experience or holidays. Relaxation and learning, enjoyment of wildlife, and proximity to nature are all pushing factors to visit national parks. The diversity of wildlife, the variety of plant and animal species, the wildness, the beauty of the scenery, and environmental safety are all pulling factors to visit national parks. In the same vein, as a result of the analysis of twelve (12) push factors, (Kim et al., 2003) identified four underlying domains including appreciation of natural resources, togetherness with the family, adventure and building social bonds and getting rid of daily routine. The most influential push factors were the appreciation of natural resources and the adventure and establishing social bonds. From the factor analysis of 12 pull factors, three main domains that were created are information and convenience of recreation facilities, tourist resources, and affordability of transport facilities. The accessibility and how close the parks are to most residential areas are key success factors for the effective management of national parks.

Methodology

The dataset used in the current study consists of 2436 monthly observations from 2009-2021 that were collected from Volcanoes National Park (918 observations), Akagera National Park (753 observations) and Nyungwe National Park (763 observations) respectively. Though different authors have attempted to use AI-based techniques to identify factors affecting visits to national parks (Clark et al, 2019; Rice et al., 2019), the main limitation of previous studies on this topic, however, is the failure to apply machines learning techniques to forecast the national park visits. The results of linear regression and random forest are therefore compared at both training set and test using Root Mean Square Error and R-square to determine the best models to predict visits to National Parks.

More recently, researchers have shown a growing interest in the use of regression techniques (de Castro et al., 2015). Regressions techniques that have been frequently used include correlation and linear regression (Morgan, 2011; Smeral, 2009); Panel data regression (Kim & Song, 2017; Meric & Hunt, 1998; Nerg, 2012); Spatial regression (Hanink &Stutts, 2002). The main determinants that were used are park characteristics and quality of products offered (de Castro et al., 2015); Schägner, 2016), demographic and behavioral characteristics of the visitors (Kim & Song, 2017; Meric & Hunt, 1998.) financial and non-financial shocks (Morgan et al., 2011; Smeral, 2009), climate conditions (Schägner et al., 2016), distance and travel costs (Hanink & White, 1999; Manfredo et al., 1996). Unlike Multivariate Adaptive Regression Splines (MARS), regression models are not complicated as they are not multiplied each other, and they do not suffer from overfitting the predictive power of the models.

Regression tree is one of the AI-based forecasting techniques that provide a considerable advantage due to the ability to readily create regression trees and their non-parametric architecture (Cankurt, 2016). The leaf nodes of a regression tree yield a single value, whereas the leaf nodes of a model tree generate regression models. When compared to regression trees, Random Forest frequently produces more compact and accurate results. At the decision tree nodes, model trees are essentially a combination of a traditional decision tree and multiple distinct linear regression function options (Wang & Witten, 1997; Solomatine & Xue, 2004). Decision Trees (DTs) are a nonparametric supervised learning technique that can be used to solve classification and regression issues. Classification and regression tree (CART) (Breiman, Friedman, Olshen & Stone, 1984), multivariate adaptive regression splines (MARS) (Friedman, 1991), and M5 (Friedman, 1991) are the most popular regression tree methods for predicting real values (Quinlan, 1992; Friedman, 1991).

Random Forest (RF) techniques have recently attracted a lot of attention from a variety of industries (Genuer et al., 2017; Hapfelmeier & Ulm, 2014). They are valued because they can process vast volumes of data to extract small groups of variables and determine variable relevance (Matin et al., 2018). RF is a popular and effective approach for classification and addressing regression issues. It is based on model aggregation ideas (Abellán et al., 2017; Breiman, 2001; Grömping, 2009) The basic idea of RF is to produce a large number of binary decision trees using two processes (Lulli et al., 2019). To begin, instead of using the complete sample, utilize bootstrap samples and the second option is to use a randomized tree predictor instead of classification and regression trees on each bootstrap sample (CARTs). For regression issues, the generating method includes averaging individual tree forecasts. The best split is then generated by selecting a random subset of the variables and searching through them. After that, the optimum split is formed by randomly selecting a subset of the variables and looking through them. Out-of-bag (OOB) values are delivered down the tree after this is repeated at each node to obtain the error rate and variable significance estimates (VI).

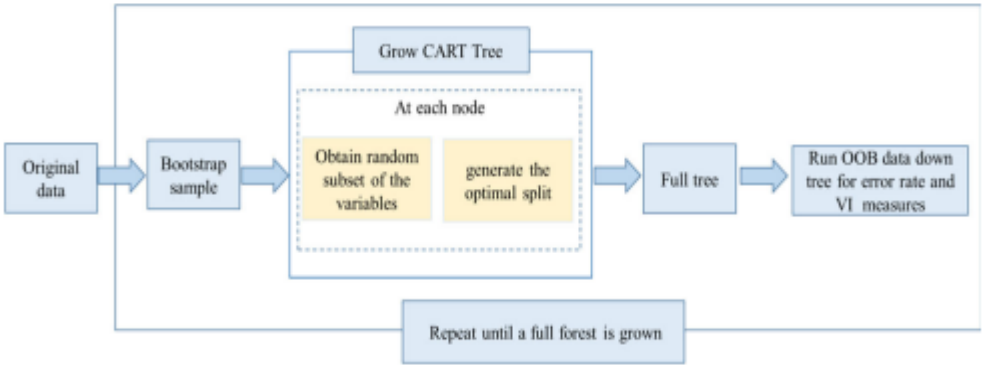

Figure 1. Random forest algorithm

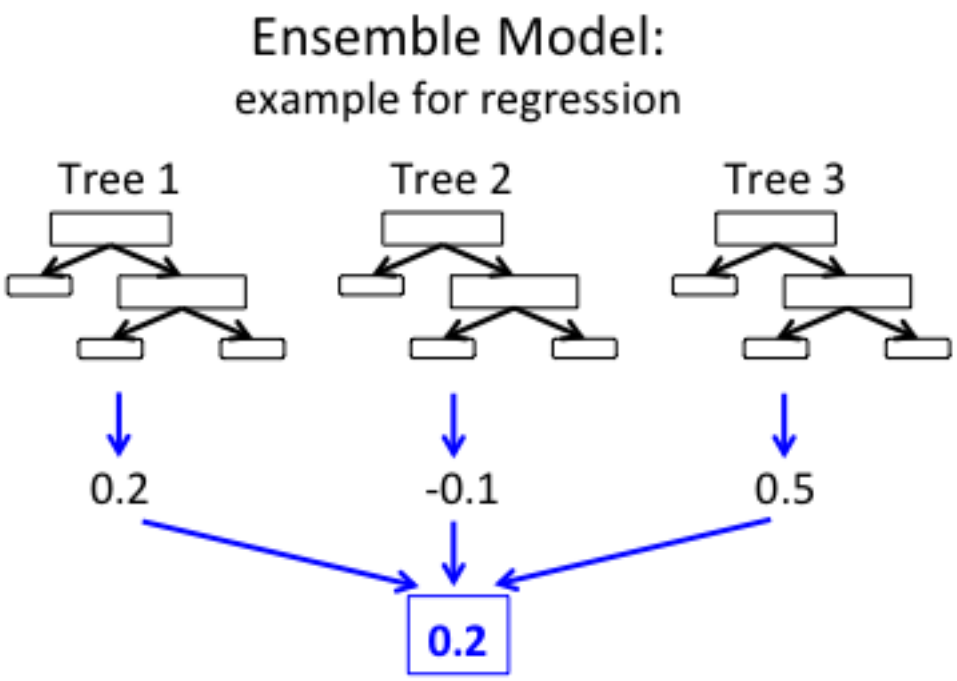

Figure 2: Random forest model

The random forest approach uses 6000 estimators, eight maximum depths, eight maximum features, four minimum sample leaves, and a random state of zero. To achieve the same scalability, a logarithm transformation was used.

Variables definition

Dependent variable: For the various models used in the current study, the number of visitors to national parks is a dependent variable.

Independent variables: The table below lists the independent variables used in the current study. These variables include the price to visit different products under National parks, dummy variables to capture the effects of different months of the year, dummy variables to capture the effects of national park characteristics (different products in national parks), trend to capture the effect of time, and a dummy variable to capture the effect of Covid-19 on visitation to national parks.

Model specification

In the current study, four models are estimated. First, we learn a model that combine together data for all three national parks and all variables and then we learn models for each National Park individually

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \cdots \ldots + \beta_n X_n$$

β_0 stands for the intercept while βs represent different explanatory variables

Y Stands for the dependent variable that is the number of visitors to national parks.

Xs represent different explanatory variables or features, which are used in different models. These variables include the price of different products, dummy variables to capture the effects of different months of the year, dummy variables to capture the effects of park characteristics (park products), trend to capture the effect of time and the dummy variable to capture the effects of Covid-19.

Table 1. description of different variables used in the models

Variable/ feature	Variable description
Price	Price of different products
Month January	"1" stands for visits in January and "0" otherwise
Month_February	"1" stands for visits in February and "0" otherwise
Month_March	"1" stands for visits in March and "0" otherwise
Month_April	"1" stands for visits in April and "0" otherwise
Month_May	"1" stands for visits in May and "0" otherwise
Month_June	"1" stands for visits in June and "0" otherwise
Month_July	"1" stands for visits in July and "0" otherwise
Month_August	"1" stands for visits in August and "0" otherwise
Month_ September	"1" stands for visits in September and "0" otherwise
Month_October	"1" stands for visits in October and "0" otherwise
Month_November	"1" stands for visits in November and "0" otherwise
Month_December	"1" stands for visits in December and "0" otherwise
trend	Trend is introduced in the model to capture the effects of time on visitation to national parks
Covid-19_False	A dummy variables that takes 1 for the period from March 2020 until now
Covid-19_True	A dummy variables that takes 0 for the period that is not from March 2020 until now
Park_Akagera	"1" stands for visit to Akagera National Park and "0" otherwise
Park_Nyungwe	"1" stands for visit to Nyungwe National Park and "0" otherwise
Park_Volcanoes	"1" stands for visit to Volcano National Park and "0" otherwise
Product_Birding	"1" stands for when going Birding is preferred by a visitor and "0" otherwise
Product_Boat ride	"1" stands when going Boat ride is preferred by a visitor and "0" otherwise
Product_Camping	"1" stands for when going for Camping is preferred by a visitor and "0" otherwise
Product_Canopy	"1" stands for when going for Canopy is preferred by a visitor and "0" otherwise
Product_Caves	"1" stands for when visiting Caves is preferred by a visitor and "0" otherwise
Product_Dian Fossey's Tomb	"1" stands for when visiting Dian Fossey's Tomb is preferred by a visitor and "0" otherwise
Product_Fishing & Other	"1" stands for when going for Fishing & Other is preferred by a visitor and "0" otherwise
Product_Game Safari	"1" stands for when Game Safari is preferred by a visitor and "0" otherwise
Product_Golden Monkey	"1" stands for when visiting Golden Monkey is preferred by a visitor and "0" otherwise
Product_Gorillas	"1" stands for when visiting Gorillas is preferred by a visitor and "0" otherwise
Product_Mountain climbing	"1" stands for when Mountain climbing is preferred by a visitor and "0" otherwise
Product_ Nature Walk	"1" stands for when Nature Walk is preferred by a visitor and "0" otherwise
Product_Primates	"1" stands for when visiting is preferred by a visitor and "0" otherwise
Product_Trails	"1" stands for when going for Trails is preferred by a visitor and "0" otherwise

Results and discussion

The major purpose of this study is to use machine learning techniques to determine factors affecting the number of visitors to national parks. The results of random forest and linear regression models on both the train and test sets are compared using Root Mean Square Error (RMSE) and R-Squared to determine the best models for predicting national park visitation. Though both random forest models and linear regression models were able to better forecast

visits to national parks on the train set, neither of them was able to do so on the test set. This can be explained by the fact that the training set contained data from before COVID-19, whereas the test set comprised data recorded during COVID-19 with different behaviors than before COVID-19. Random forest models outperformed linear regression models in predicting the number of visitors to national parks. Akagera National Park outperformed Volcanoes and Nyungwe National Parks, notably on the train set. The random forest, therefore, outperforms other models in predicting the number of visitors to national parks. The most important factors in determining the number of visits to national parks are price, park-specific qualities, and different months of the year, each of which has a distinct significance depending on the park. Though the explanatory power of covid-19 has been relatively small (0.04%), The results show that Covid-19 had some impact on the number of visitors to National Parks.

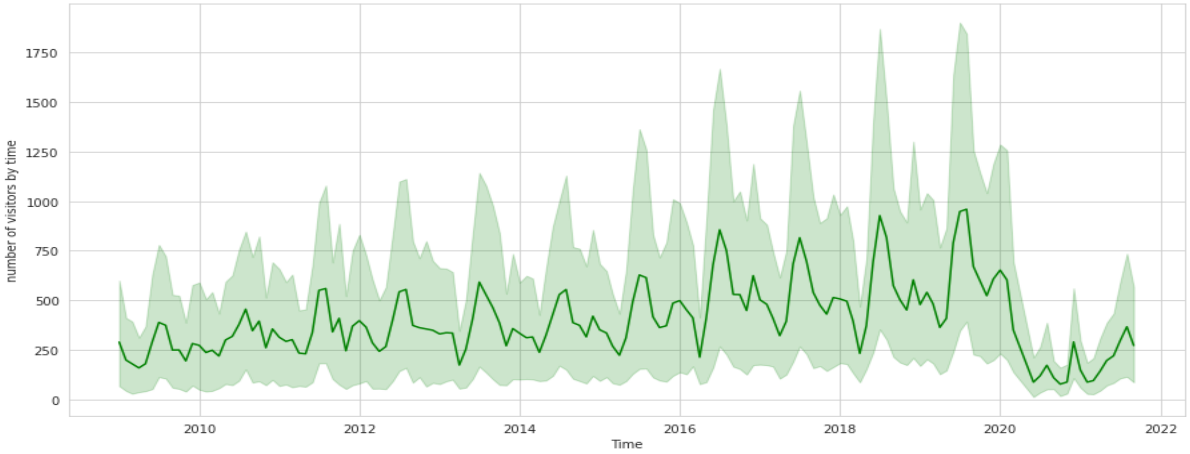

Figure 2. Trends in Visitation to National Parks

Figure 2 depicts the trend in visits to national parks overtime. As the graph shows, the number of visitors to national parks has been fluctuating overtime. Most importantly, during COVID-19, the number of visitors sharply declined.

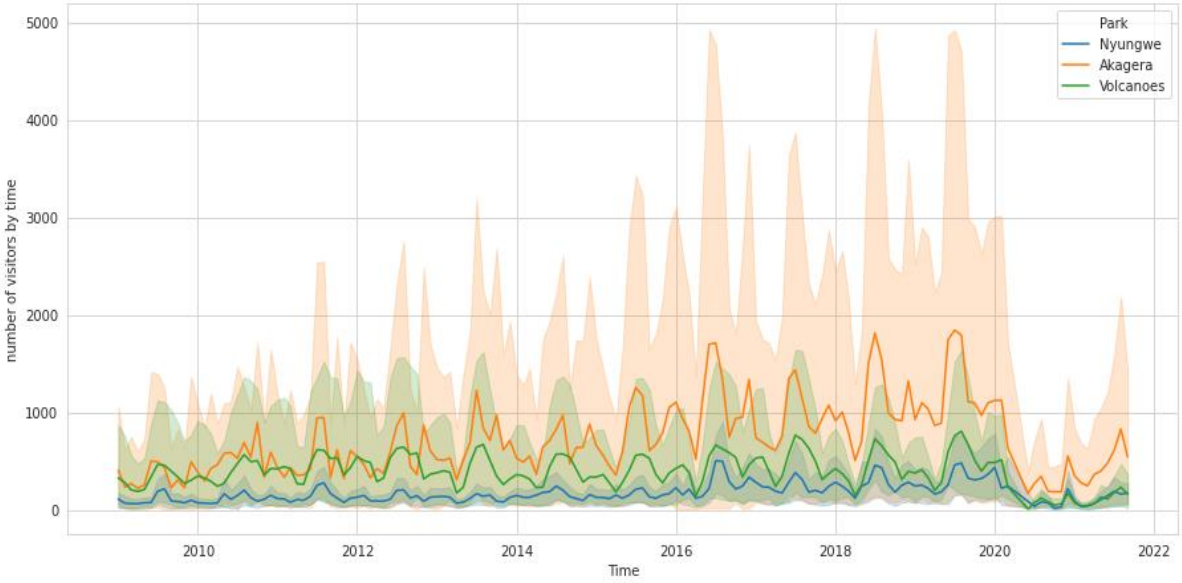

Figure 3. Comparison of overtime variation in visiting different products in National Parks

Figure 3 depicts an overtime variation in the demand for products in Akagera, Volcanoes, and Nyungwe National Parks. Overall, the demand for products in Akagera National Park has been volatile, particularly in the aftermath of the COVID-19 outbreak.

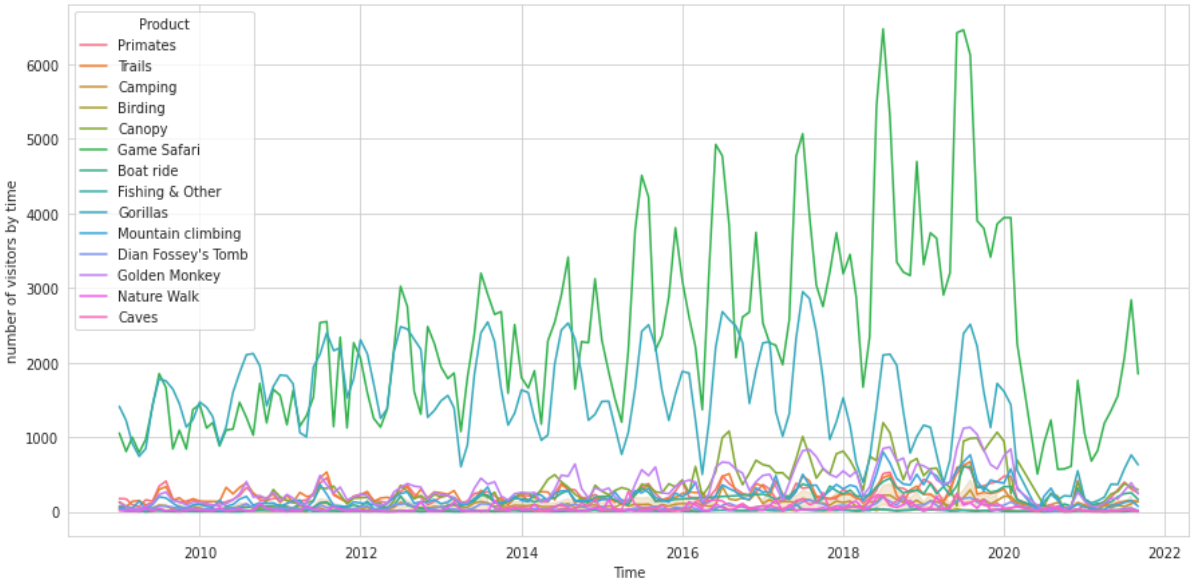

Figure 4. Overtime variation in the demand for different products in National Parks

The Figure 4. Illustrates demand for various products in national parks . Though the demand for products such as Game Safaris has consistently been higher than other products, demand for all products has fluctuated over time, with Game Safaris and Gorillas dominating others, particularly in the months following the COVID-19 pandemic outbreak. The Figure 5 below illustrates overtime monthly variation in the number of visitors to national parks. From the graph it can be seen that the number of visitors across all National Parks fluctuates throughout the year. Moreover, Though the number of visitors to Akagera National Park appears to be higher than Volcanoes and Nyungwe National Parks, the demand tends to be high in summer (from June to August) and at the end and beginning of the year. Similarly, all national parks are more likely to see a marked decline in the number of visitors in April.

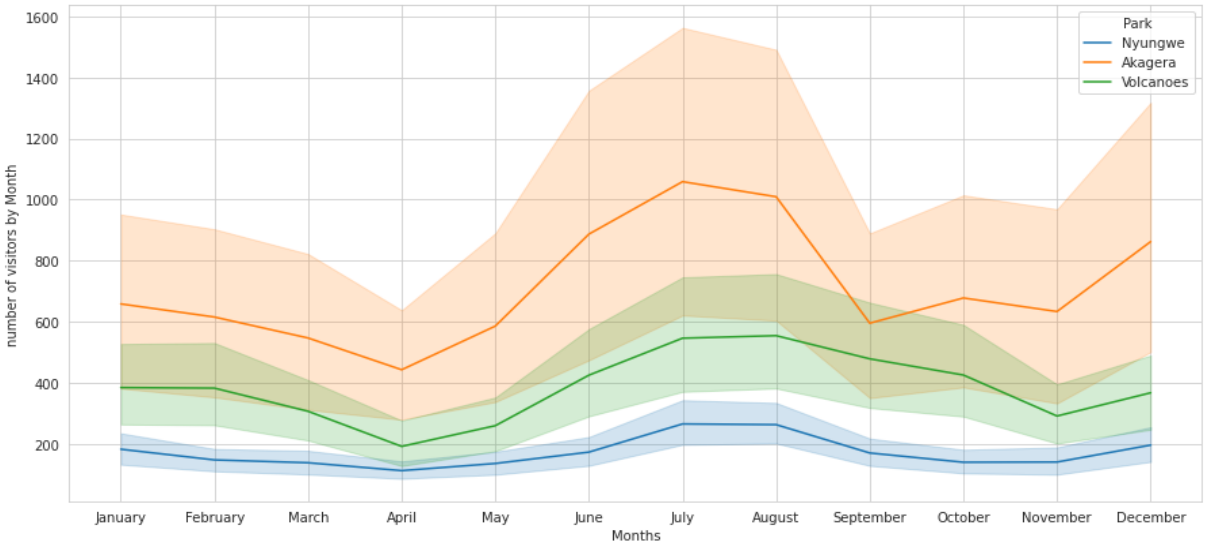

Figure 5. Comparison of monthly variations in the number of visitors to National Parks

Model evaluation for all national parks

To determine the extent to which each model can accurately estimate the visitation of the concerned National Park, it is critical to assess the predictive power for each model. This is done by comparing the true values to the projected values of the model. The generalizability of the model to new data must also be evaluated, which is performed by splitting the dataset into two sets, one for training and the other to test the generalizability of the model. In the current study, Root Mean Square Error (RMSE) and R-square (R2) are used. Both measures are the most commonly used approaches for evaluating model accuracy. RMSE: The Root Mean Square Error (RMSE) is a measure of the residuals of the standard deviation (prediction errors). The errors are squared before being averaged with RMSE. Residuals are referred to as the difference between expected and actual values. The RMSE is a measure of the distribution of residuals. When forecasting errors are severe, the root mean square error (RMSE) plays an important role to reduce them.

$\boldsymbol{R^2}$: The R-square (R^2) is a statistical measure of how much variance is accounted for in a relationship between two (or more) variables. This metric shows how well a model fits the data and how closely projected values match actual values. The R squared value varies from 0 to 1, with 0 indicating that the model does not match the data and 1 indicating that the model fits the data well (100 percent).

Table 2: The results of model evaluation for all National Parks combined

Park	Model	Set	RMSE	R-square
All parks	Linear Regression	Train	0.580	0.884
		Test	1.187	0.565
	Random Forest	Train	0.536	0.901
		Test	1.179	0.571

Table 2 presents the results of the model evaluation for Akagera, Volcanoes and Nyungwe National Parks. The root mean squared error of linear regression is 0.580 on a training set and 1.187 on the test set. R-square of linear regression model is estimated at 0.884 On a training set and 0.565 on a test set. The linear regression model can therefore accurately predict the number of visitors to National Parks with an accuracy of 88.4 percent on the train set and 56.5 percent on the test set given a month of visit and the national park visited. In summary, the regression model performed better on the train set, which contained data prior to the outbreak of COVID-19, but not on the test set (56.5 percent), which comprised data and behaviors during the COVID-19 period, during which the number of visitors to the national park dropped substantially.

The results in Table 2 indicate that root mean squared error for random forest model is 0.536 on a train set and 1.179 on a test set. The R-square of random forest is 0.901 on a train set and 0.571 on the test set. The model can therefore better predict the number of visitors with 90.1 percent accuracy on the train set and 57.1 percent accuracy on the test set. The random forest model therefore made better predictions on its training set, which contained data before COVID-19, but it failed on the test (57.1 percent) due to the fact that the test set contained different data and different behaviors compared to the period before COVID-19.

Table 3: The results of model evaluation for Akagera National Park

Park	Model	Set	RMSE	R-square
Akagera	Linear Regression	Train	0.505	0.937
		Test	0.938	0.648
	Random Forest	Train	0.450	0.950
		Test	1.04	0.760

Table 3 depicts the results of the model evaluation for Akagera National Park. The root mean squared error of linear regression On a training set is 0.505 and 1.33 on a test set. In the same vein, the R-square of linear regression is 0.868 on a training set and 0.648 on a test set. The linear regression model can therefore accurately predict the number of visitors to Akagera National Park with an accuracy of 93.7 percent and 64.8 percent on both train and test sets. Therefore, the model made better predictions on its training set, which contained data before Covid-19, but it failed to make better predictions on a test (64.8 percent) because during the Covid-19 period, the test set contained different data and different behaviors compared to the period before Covid-19.

The results in the Table 3 above also indicate that the root mean squared error of random forest model is on 0.450 a training set and 1.04 on a test set. Random forest has an R-square of 0.950 on a train set and 0.760 on a test set. The model can therefore accurately forecast the number of visitors to Akagera National Park with 95.0 percent accuracy on the train set and 76.0 percent on the test set. Taken together, the model made better predictions on its training set, which consisted of data before COVID-19. Though the test set was mostly made of data from the COVID-19 period, random forest models made better predictions with an R-square estimated at 76 percent.

Table 4: The results of model evaluation for Volcanoes National Park

Park	**Model**	**Set**	**RMSE**	**R-square**
Volcanoes	Linear Regression	Train	0.577	0.868
		Test	1.33	0.359
	Random Forest	Train	0.509	0.89
		Test	1.336	0.354

The results of the model evaluation for Volcanoes National Parks are shown in Table 4 above. The root mean squared error of linear regression is 0.577 on a training set and 0.938 on a test set. The R-square of linear regression is 0.868 on a training set and 0.359 on a test set. The linear regression model can accurately predict the number of visitors to Volcanoes National park at a level of 86.8 percent and 35.9 percent on the test set. As a result, the model predicted better on its training set, which consisted of data prior to COVID-19, but failed to predict better on a test because the test set contained different data and different behaviors compared to the period before Covid-19.

As indicated by the results in the Table 4 above, the root mean squared error of the random forest model is 0.509 on a training set and 1.336 on a test set. R-square of the random forest model is 0.89 on a train set and 0.354 on a test set. The model can therefore predict the number of visitors to Volcanoes National Park with an accuracy of 89.0 percent on the training set and 35.4 percent on the test set. Taken together, the model made better predictions on its training set, which was made of data before COVID-19, but it failed on a test set because the test set contained different data and different behaviors compared to the period before Covid-19. Both random forest and regression models performed better on train set than test set.

Table 5: The results of model evaluation for Nyungwe National Park

Park	**Model**	**Set**	**RMSE**	**R-square**
Nyungwe	Linear Regression	Train	0.583	0.845
		Test	1.157	0.477
	Random Forest	Train	0.488	0.891
		Test	1.133	0.500

Table 5 presents the results of the model evaluation for Nyungwe National Park. The root mean squared error of linear regression is 0.583, and 1.157 on a test set. The R-square of linear regression on a training set is 0.845 and 0.477 on a test set. The linear regression model can

accurately predict the number of tourist visitors given a month of visit and product with an accuracy of 84.5 percent on the train set 47 percent on the test set. Therefore, the model made better predictions on its training set, which was made of data before COVID-19, but it failed to make better predictions on a test because the test set contained different data and different behaviors compared to the period before COVID-19.

As the results in the Table 5 indicate, the root mean squared error of the random forest model is 0.488 On a training set and 1.133 on the test. Random forest has an R-square of 0.891 on a train set and 0.50 on a test set. The model can predict the number of visitors given to Nyungwe National Park with an accuracy of 89.1 percent on the training set and 50.0 percent on the test set. In summary, the model made better predictions on its training set, which was made of data before COVID-19, but it failed to make better predictions on a test set because the test set contained different data and different behaviors compared to the period before COVID-19. Taken together, both random forest and regression models made better predictions on training, but all failed to make better predictions on the test set. Both linear regression and random forest models performed better using data from Akagera National Park than Volcanoes and Nyungwe National Parks.

Factors affecting the number of visitors to National Parks

The explanatory power of variables to model prediction is indicated by feature importance. It shows the influence of various variables in estimating the number of visitors to national parks in our case.

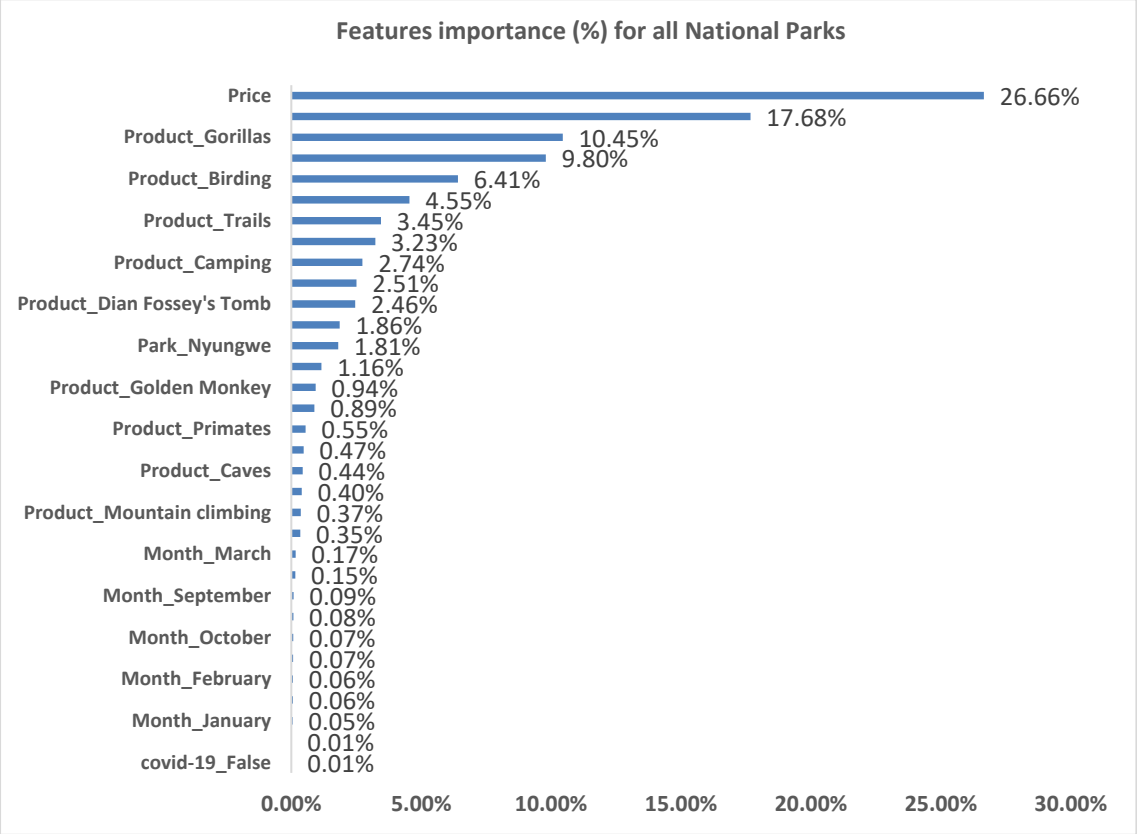

Figure 6: Feature importance for all National Parks using Random Forest

The most important features to explain the number of visits to national parks, as shown in Figure 6 above, are the price (26.6 percent), visiting game safaris (17.6%), gorillas (10.4 percent), fishing and other (9.8%), birding (6.4 percent), the trend (4.5 percent) that captures

the effect of time on visits to national parks; trails (3.4 percent); nature walks (3.2 percent); camping (3.2 percent); and camping (3.2 percent). Though the feature importance for different months of the year is relatively small compared to other factors, July (0.47%), August (0.4%), and April (0.35%) appear to have more influence than other months of the year. Although the coefficient of COVID-19 is relatively smaller than other features, it has had some impact on visitors to national parks as indicated in Figure 6.

Factors affecting the number of visitors to Akagera National Park

Figure 7 below depicts the explanatory power of different factors in predicting the number of visitors to Akagera National Park. The most important factors that explain visit to Akagera National Park are Game Safari (42.7 %); fishing and others (27.5%); price (12.5%), trend that capture the effect of time on visits to Akagera National Park (5.9%); camping (5.2%); boat ride (4.4%) ; camping (5.2%) and boat ride (4.5%). In the same vein, the months of the year that are more important than others to explain visits to Akagera National Park are July (0.3%), August (0.2%), and March (0.17%) and April (0.13%).

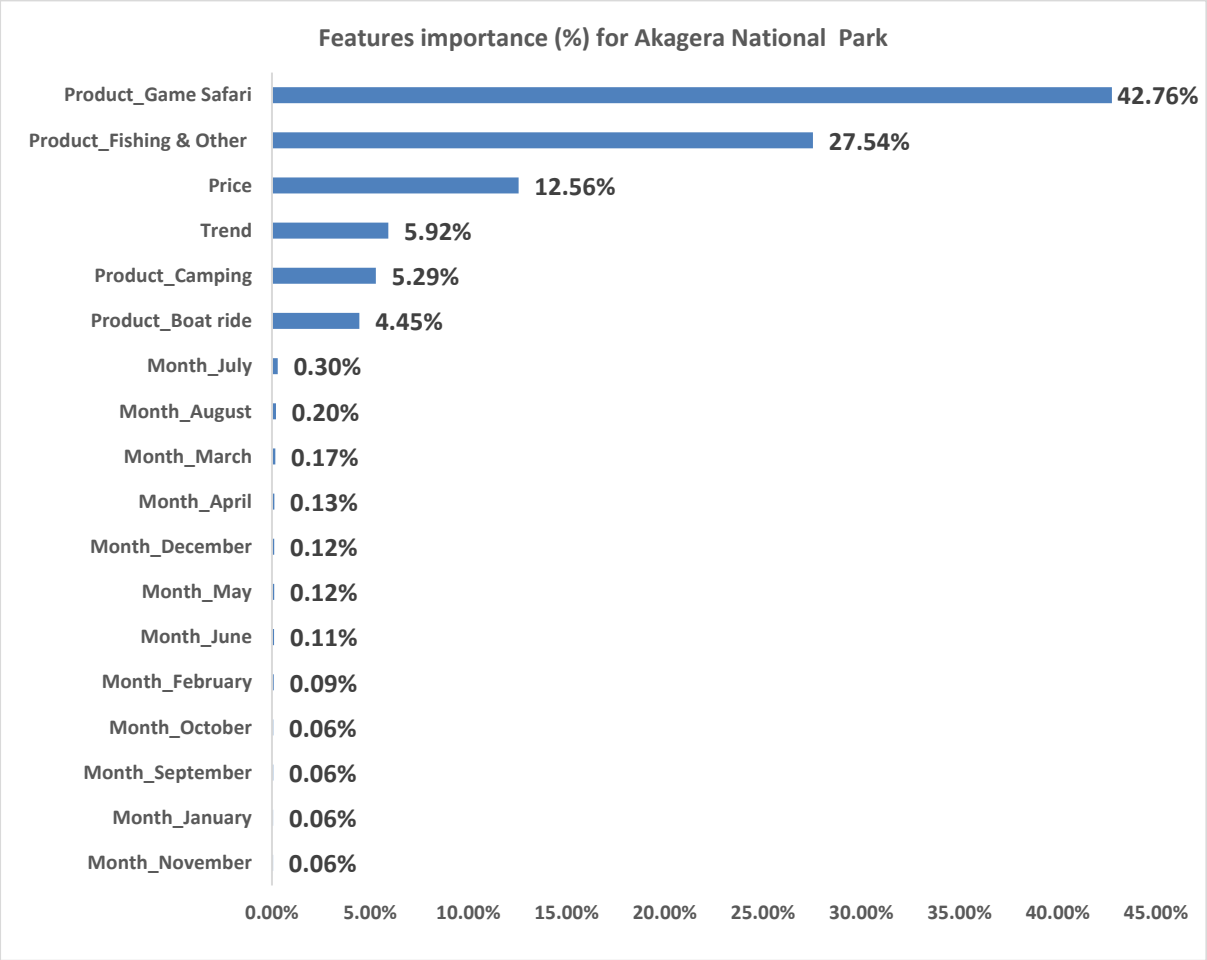

Figure 7: Features importance for Akagera National Park using Random Forest.

Factors affecting the number of visitors to Volcanoes National Park

Figure 8 illustrates the importance of variables in predicting visits to Volcanoes National Park. The most important factors to explain visits to Volcanoes National Park include the price to visit different products (40.2%); visiting gorillas (26.5%); nature walk (12.1%); trend (6%); Dian Fossy's tomb (3.7%); Golden Monkey (3%) and mountain climbing (2.5%). In addition,

the most important months of the year to explain the number of visitors to Volcanoes National Park are April (1%), July (0.69%), August (0.57%) and May (0.41%).

As illustrated by Figure 8, the effect of COVID-19 on the number of visitors to Volcanoes National Park has been relatively small (0.04%).

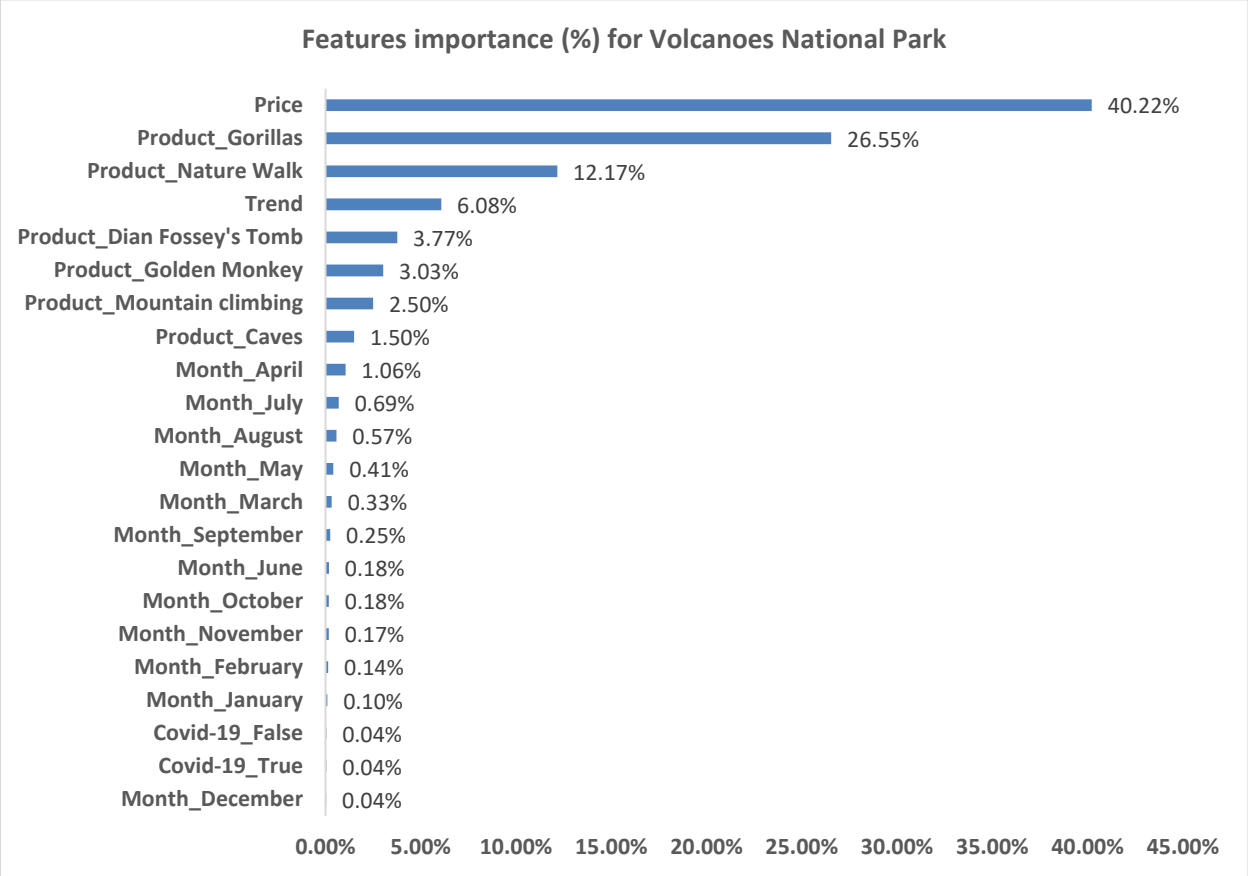

Figure 8: Feature importance for Volcanoes National Park using Random Forest

Factors affecting the number of visitors to Nyungwe National Park

Figure 9 above presents the importance of different explanatory variables to predict the number of visitors to Nyungwe National Park. From the results of random forest model, the most important factors to explain the visiting to Nyungwe National Park include price (30.16%), birding (23.7%); camping (13.7%); trails (12.1%); trend (6.3%); canopy (5.9%) and primate (3.8%). Three months of the year that are more likely to have higher influence than others are August (1.1.8%); July (1.14%) and March (0.31%).

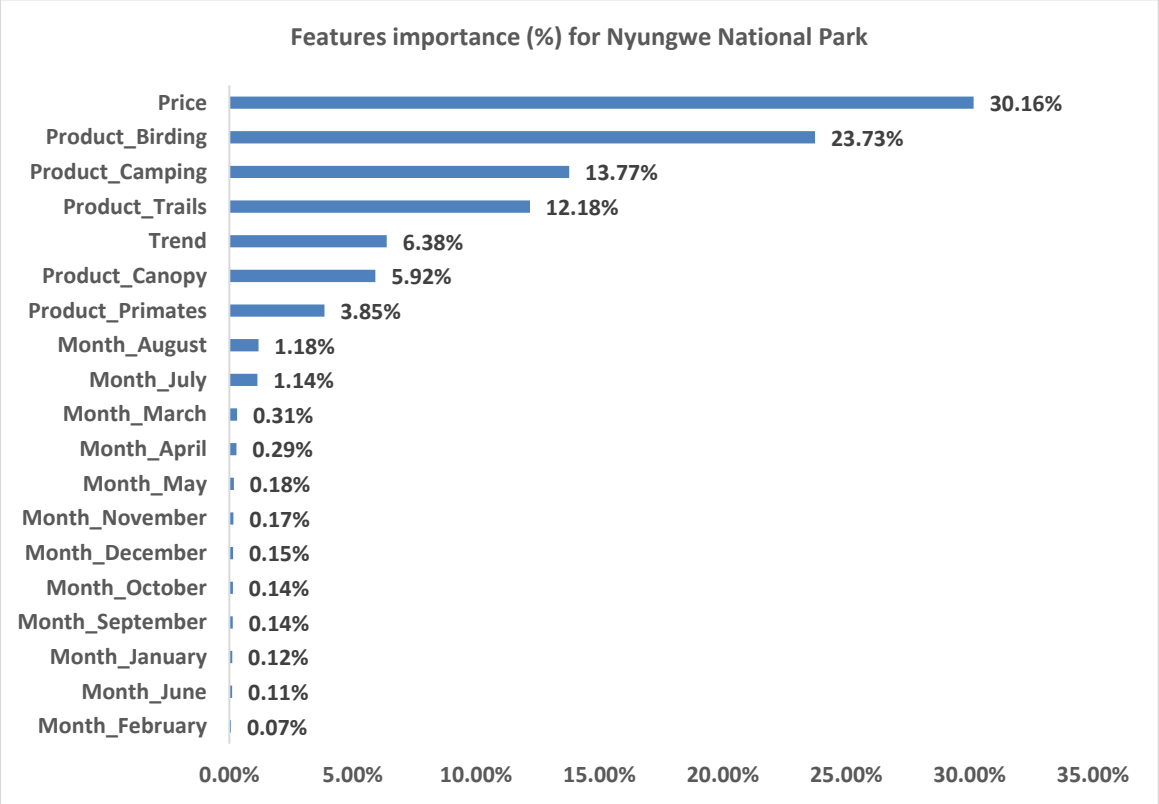

Figure 9: Feature importance for Nyungwe National Park using Random Forest

Conclusions

Taken together, both random forest models and linear regression models were able to predict the visits to national parks better on the train set, but all failed to make better predictions on the test set. This can be explained by the fact that the training set contained data prior to COVID-19, while the test set consisted of data during the COVID-19 period with different behaviors compared to the period before COVID-19. Random forest models outperformed linear regression in forecasting visitors to national parks. The results of Akagera National Park outperformed those of Volcanoes National Park and Nyungwe National Park, notably on the train set. Overall, the random forest outperforms other regression models in predicting the number of visitors to national parks. The most important factors to explain the number of visits to national parks are price of different products in National Parks, park-specific characteristics, and different months of the year, whose relevance varies from one park to another.

The number of visitors tends to vary from one park to another and the demand for different products tends to be seasonal, with the demand being higher in summer (from June to August) and at the end and beginning of the year, in December and January. Similarly, all national parks are more likely to see a marked decline in the number of visitors in April. Given that national parks and other nature-protected areas are primary sources of revenue for the government, and that the supply of various products in national parks remains relatively constant, it is necessary to identify drivers of demand for products under National Parks in order to make more accurate predictions revenues from visiting National Parks. It also allows for the provision of accurate and timely information, as well as efficient booking windows and staffing. According to the current study, random forest model appears to perform better in predicting visitors to national parks. The current study, therefore, suggests that random forest proves to be the best model for forecasting visits to national parks. The findings from this study also highlight the seasonality nature of the demand for products under National Parks, which should be taken into account while projecting revenue from visiting National Parks.Future

research should explore the possibility to include more parameters in the model to measure the relative demand for products in National Parks. The effects of climate, level of income of the visitor and promotion expenditure as significant predictors can be evaluated, thus filling a long-standing gap in the literature related to visits to National Park. When exploring RDB database about national parks data, this study found a variety of limitations. The use of monthly and aggregated data did know allow access to some key visitor characteristics such as gender, age, country of origin, income of the visitor, education level of the visitor, the length of stay, which would have been used to improve the model prediction. Therefore, Future studies need to explore the possibility of using daily data and include the aforementioned metrics to improve the model accuracy.

The number of visitors seemed to vary from one national park to another. Therefore, future research can deeply investigate the extent to which individual characteristics of national parks can determine the number of visitors to that specific national park. Finally, our successful application of machine learning techniques to predict the number of visitors to national parks offers researchers the possibility of applying these approaches to other aspects of national parks that are increasingly attracting the interest of researchers. Therefore, an attempt can be made to predict revenues from visiting National parks, and the demand for various products in National Parks. This would assist park management with effective planning, allocation and use of resources from visiting National Parks.

Acknowledgements
My special thanks go to the African Centre of Excellence in Data Science-ACEDS, the University of Rwanda for having offered financial support. I also extend my sincere appreciation to my PhD Supervisors, Prof. Herman Musahara and Dr. Edouard Musabanganji for their continuous support throughout my research journey.

References

Abellán, J., Mantas, C. J. & Castellano, J. G. (2017). A Random Forest Approach Using Imprecise Probabilities. *Knowledge-Based Systems*, 134, 72-84.

Breiman, L. (2001). Random Forests. *Machine learning*, 45(1), 5-32.

Breiman,L., Friedman, J. H. Olshen, R. A. & Stone, C. (1984). Classification and Regression Trees. *Monterey, Calif., U.S.A.: Wadsworth, Inc.*

Cankurt, S. (2016). Tourism demand Forecasting Using Ensembles of regression Trees. *International Conference on Intelligent systems*,702-708.

Clark, M., Wilkins, E.J., Dagan, D.T., Powell, R., Sharp, R.L. & Hillis, V. (2019). Bringing Forecasting into the Future: Using Google to Predict Visitation in US National Parks. *Journal of Environmental Management*, 243, 88-94.

de Castro, E.V., Souza, T.B. & Thapa, B. (2015). Determinants of Tourism Attractiveness in the National Parks of Brazil Parks. 21(2), 51-62.

Fredman, P., Friberg, L.H. & Emmelin, L. (2007). Increased Visitation from National Park Designation. *Current Issues in Tourism*, 10(1), 87-95.

Genuer, R., Poggi, J. M., Tuleau-Malot, C. & Villa-Vialaneix, N. (2017). Random Forests for Big Data. *Big Data Research*, 9, 28-46.

Grömping, U. (2009). Variable Importance Assessment in Regression: Linear Regression Versus Random Forest. *The American Statistician*, 63(4), 308-319.

Halpenny, E. (2010). Pro-environmental Behaviours and Park Visitors: The Effect of Place Attachment. *Journal of Environmental Psychology, 30(4)*, 409-421.

Hanink, D.M. & Stutts, M. (2002). Spatial Demand for National Battlefield Parks. *Annals of Tourism Research*, 29(3), 707-719.

Hanink, D.M. & White, K. (1999). Distance Effects in the Demand for Wildland Recreational Services: The Case of National Parks in the United States. *Environment and Planning*, 31(3), 477-492.

Kim, H. & Song, H. (2017). Measuring Hiking Specialization and Identification of Latent Profiles Of Hikers. *Landscape and Ecological Engineering*, 13(1), 59-68.

Kim, S.S., Lee, C.K. & Klenosky, D.B. (2003). The Influence of Push and Pull Factors at Korean National Parks. *Tourism Management*, 24(2),169-180.

Kruger, M. & Saayman, M. (2015). The Determinants of Visitor Length of Stay at the Kruger National Park. Koedoe: *African Protected Area Conservation and Science*, 56(2),1-11.

Kruger, M. & Saayman, M. (2010). Travel Motivation of Tourists to Kruger and Tsitsikamma National Parks: A comparative study. *South African Journal of Wildlife Research*, 40(1), 93-102.

Leung, Y.F., Spenceley, A., Hvenegaard, G., Buckley, R. & Groves. (2018). *Tourism and Visitor Management in Protected Areas : Guidelines for sustainability (Vol 27).* Gland, Switzerland: IUCN.

Lulli, A., Oneto, L. & Anguita, D. (2019). Mining Big Data with Random Forests. *Cognitive Computation*, 11(2), 294-316.

Manfredo, M.J., Driver, B.L. & Tarrant, M.A. (1996). Measuring Leisure Motivation: A Meta-Analysis of the Recreation Experience Preference Scales. *Journal of Leisure Research*, 28(3), 188-213.

Manning, R. (1980). International Aspects of National Park Systems. *International aspects of national park systems: Focus on tourism*, 179-192.

Manning, R. (2011). Indicators and Standards in Parks and Outdoor Recreation. In Quality-of-Life community indicators for parks. *Recreation and Tourism Management*, 11-22.

Matin, S. S., Farahzadi, L., Makaremi, S., Chelgani, S. C. & Sattari, G. H. (2018). Variable Selection and Prediction of Uniaxial Compressive Strength and Modulus of Elasticity by Random Forest. *Applied Soft Computing*, 70, 980-987.

Meric, H.J.& Hunt, J. (1998.). Ecotourists' Motivational and Demographic Characteristics: A Case of North Carolina Travelers . *Journal of Travel Research*, 36(4), 57-61.

Morgan, K.L., Larkin, S.L. & Adams, C.M. (2011). Empirical analysis of Media Versus Environmental Impacts on Park Attendance. *Tourism Management*, 32(4), 852-859.

Mutanga, C.N., Vengesayi, S., Chikuta, O., Muboko, N. & Gandiwa, E. (2017). Travel Motivation and Tourist Satisfaction with Wildlife Tourism Experiences in Gonarezhou and Matusadona National Parks, Zimbabwe. *Journal of Outdoor recreation and tourism*, 20, 1-18.

Nerg, A., Uusivuori, J., Mikkola, J., Neuvonen, M. & Sievänen, T. (2012). Visits to national parks and hiking areas: a panel data analysis of their socio-demographic, economic and site quality determinants. *Tourism Economics*, 18(1), 77-93.

Neuvonen, M., Pouta, E., Puustinen, J. & Sievänen, T. (2010). Visits to National Parks: Effects of Park Characteristics and Spatial Demand. *Journal for Nature Conservation*, 18(3), 224-229.

Poudyal, N.C., Paudel, B. & Tarrant, M.A. (2013). A Time-series Analysis of the Impact of Recession on National Park Visitation in the United States. *Tourism Management*, 35, 181-189.

Quinlan, R. J. (1992). Learning with Continuous Class. *5th Australian Joint Conference on Artificial Intelligence.* Singapore.

Reihanian, A., Hin, T.W., Kahrom, E., Binti Mahmood, N.Z.& Bagherpour Porshokouh, A. (2015). An Examination of the Effects of Push and Pull Factors on Iranian national parks: Boujagh National Park, Iran. *Caspian Journal of Environmental Sciences*, 13(3).

Rice, W.L., Park, S.Y., Pan, B. & Newman, P. (2019). Forecasting campground Demand in US National Parks. *Annals of Tourism Research*, 75,424-438.

Schägner, J.P., Brander, L., Maes, J., Paracchini, M.L. & Hartje, V. (2016). Mapping Recreational Visits and Values of European National Parks by Combining Statistical Modelling and Unit Value Transfer. *Journal for Nature Conservation*, 31,71-84.

Scholtz, M., Kruger, M. & Saayman, M. (2015). Determinants of Visitor Length of Stay at Three Coastal National Parks in South Africa. *Journal of Ecotourism*, 14(1), 21-47.

Schwartz, Z. & Lin, L.C. (2006). The Impact of Fees on Visitation of National Parks. *Tourism Management*, 27(6), 1386-1396.

Smeral, E. (2009). The Impact of the Financial and Economic Crisis on European Tourism. *Journal of Travel Research*, 48(1), 3-13.

Solomatine, D.P. & Xue, Y. (2004). M5 Model Trees and Neural Networks: Application to Flood Forecasting in the Upper Reach of the Huai River in China. *Journal of Hydrologic Engineering*, 9(6), 491-501.

Thomas, C.C. & Koontz, L. (2021). Economic Effects Assessment Approaches: US National Parks Approach. In *Handbook for Sustainable Tourism Practitioners.* Edward Elgar Publishing.

Wang, Y. & Witten, I.H. (1997). Induction of Model Trees for Predicting Continuous Classes. *9th European Conference on Machine Learning.* .

Wang, Y., Wu, C., Wang, F., Sun, Q., Wang, X. & Guo, S. (2021). Comprehensive Evaluation and Prediction of Tourism Ecological Security in Droughty Area National Parks—a Case Study of Qilian Mountain of Zhangye section, China. *Environmental Science and Pollution Research*, 28(13), 16816-16829.

Zydroń, A., Szoszkiewicz, K. & Chwiałkowski, C. (2021). Valuing Protected Areas: Socioeconomic Determinants of the Willingness to Pay for the National Park. *Sustainability*, 13(2), 765.