A COMPARISON OF MACHINE LEARNING METHODS IN TOURISM

Luis Ángel Calvo Pascual
Universidad Pontificia de Comillas
lacalvo@icai.comillas.edu

Eduardo C. Garrido-Merchán
Universidad Pontificia de Comillas,
ecgarrido@icade.comillas.edu

Keywords:

Tourism, Machine Learning, Gaussian Processes, Ensemble method

Main area

Tourism

Secondary area

Quantitative methods applied to economy

A COMPARISON OF MACHINE LEARNING METHODS IN TOURISM

1. Introduction

Machine learning is the field that studies a class of methods that build functions to predict an endogenous variable as a function of other explanatory variables. In order to do so, these methods process a dataset of previous observations (Bishop, 2006). Supervised learning is the area of machine learning where the dataset of previous observations consists on a matrix of covariates or features about the variable to be predicted and an array of values that represents the variable that we want to predict. For example, consider the occupation percentage of a hotel in a particular month. A machine learning model may predict the occupation of this hotel in the future month if it is fixed with a dataset of previous occupations and different data associated with these occupations like the particular month, the mean temperature, number of sunny days, mean of the rooms price or other useful information. Most importantly, machine learning has been applied with success in a wide plethora of different disciplines such as computer science (Murphy, 2012), astrophysics (VanderPlas, 2012), renewable energies (Cornejo-Bueno, 2018) or even gastronomy (Córdoba, 2018) in the recent years.

Since machine learning techniques have revolutionized several study fields, we hypothesize that its effect in the tourism industry, both in public and private institutions, could also be very significant. Driven by that motivation, we are interested in verifying whether machine learning is being applied in the tourism industry, study the possible problems where machine learning can be applied and add empirical evidence about its usefulness with toy problems. We believe that in the tourism context, for example, machine learning, and in particular supervised learning, could be used for regression tasks, where the variable to predict is real-valued and for classification tasks, where the variable to predict is categorical-valued. As examples of regression tasks, we could predict the occupation percentage of a hotel in a particular day or month or the benefits that opening a tourist office may bring to the city hall. As examples of classification tasks, we could use it to classify whether a marketing campaign of a touristic place has caused a good or bad impression at social networks or whether it is a good idea or not to open a new hotel in a particular location. In all these problems, we are interested not only in predictions but also in the uncertainty of these predictions. Concretely, it is not the same to predict a high amount of benefits with certainty that with high uncertainty. Quantifying the uncertainty is critical for strategic decision processes and machine learning algorithms quantify uncertainty in their predictions.

The paper is organized as follows: first, we perform a brief review of the state of the art of the application of machine learning in tourism. Then, we give an overview about the fundamentals of the machine learning algorithms that we will use for the empirical experiments that are presented in a new section. As the audience of this paper is related with the tourism field, we do not provide technical details of them but the intuitions of their behaviour. We conclude this manuscript with a list of conclusions and further lines of research.

2. Machine learning in tourism

We begin this section with a short overview of the application of machine learning methods in tourism. In particular, we observe that the field that has been more targeted by these methods is the one of recommendation systems. In particular, we wish to know which is the best hotel based on my needs from a dataset of hotels. Recommendation systems in tourism use machine learning methods to predict the score of a certain hotel based on their features and in our preferences and characteristics (Nilashi, 2017). Moreover, hotel online reviews of users can also be processed by natural language techniques and machine learning systems to predict how different decisions may impact in the score that users give to certain characteristics of the hotel: hospitality, cleaning, noise... (Le, 2021). Furthermore, machine learning has also been used to forecast the demand of Chinese cruises (Xie, 2021), sentiment analysis of tourism experiences (Kirilenko, 2018) or forecasting tourist arrivals (Sun, 2019).

Despite the fact that the list includes many interesting applications of machine learning in tourism, from our point of view there is still an enourmous potential of machine learning in this field. Consequently, we will provide in Section 3 some experiments that add empirical evidence to support this claim. But first, we gently introduce a non-technical description of some machine learning algorithms that we use for our experiments to enhance the understanding of this field for the reader that is not familiarized with machine learning but wants to know what do these algorithms provide and why they are interesting.

2.1 Machine learning methods

In this work, we have explored the application of supervised learning, a particular type of machine learning problems (Murphy, 2012). We leave for future work the application of other machine learning scenarios such as unsupervised or active learning. In particular, supervised learning focuses on predicting a target **y** from observation features **X** from a dataset of observations D={**X**,**y**} where both the features **X** and the variable to predict **y** are present.

In supervised learning, our main task is to train a machine learning algorithm that approximates the theoretical underlying function f: **X** -> **y**, y=f(**x**), that explains the variable that we want to predict **y,** from the other explanatory variables **X**. In particular, if **y** is a real number, then we perform regression of **y**. On the other hand, if **y** is a categorical-valued variable, we perform classification of **y**. Moreover, in real case scenarios, **y** is usually corrupted with noise, that is, in several situations we may obtain a different y value for the same **x**. For example, we may have a particular number of costumers, variable to predict, for a given situation one year, where the situation is coded by the features that describe it such as weather, macroeconomic data or month of the year, and a different number of costumers of the same situation the following year. Hence, we are interesting in approximating the general trend y=f(**x**) that, in average, happens through a large number of years. More precisely, we define the generalization error *e* as the amount of error that we incur on when we predict this trend over the years in average *e* = E(|**y**-f(**x**)|), where E is the expectation of the absolute difference between the predictions of the model f(**x**) and what it really happens **y**. Machine learning algorithms minimize the estimation of the *e* error parameter via different methodologies as k-fold cross validation or train test splitting. The idea is that we want to find an estimator of the error with low bias and variance, that is, our predictions have low bias and variance. If bias is high, we incur in underfitting, our model is not complex enough to approximate reality and if variance is high we incur in overfitting, that is, our model is fixing noise and not the general trend. To sum up, machine (supervised ??) learning models generalize the dataset of previous observations D={**X**,**y**} to approximate a function y=f(**x**) that captures the trend between the explanatory variables **x** and the variable that we are trying to predict y.

Having reviewed the basics of machine learning, we now briefly provide the most fundamental details of some of the algorithms that we test on the experiments reported on Section 3.

First, one of the most basic machine learning methods whose fundamental concepts come from the statistics field in linear regression (Bishop, 2006). In simple linear regression, we have as many **β** real-valued parameters as number of explanatory variables plus one, called the intercept. These parameters are obtained from an analytical expression via the ordinary least squares method. In particular, each **β** parameter represents how does the predicted variable change when the associated explanatory variable of the parameter changes, *ceteris paribus*. Concretely, the method assumes that the predicted variable is explained by a linear combination of all the explanatory variables. If the task to perform is classification, logistic regression squashes the output of linear regression in a sigmoid function in the task of binary classification and a softmax function in the case of multi-class classification, giving a probability distribution of the class. Both methods, although being simple and interpretable, lack of the predictive performance of more sophisticated methods such as Gaussian processes or random forests, that are the two methods that have better predictive performance in our experiments.

Gaussian processes are a generalization of the multivariate Gaussian distribution N(μ, ${}_{\Sigma}$) that represents a distribution over functions (Rasmussen, 2003). In particular, we assume that the generalization error of the machine learning algorithm is one of those functions. Gaussian processes are non-parametric methods that output a predictive distribution of the explanatory variables **X** at each of its possible values **X**=**x**. Then, they provide a prediction ***y*** for each of the observations **x** and an uncertainty *σ* over that prediction. The joint probability distribution of those predictions is the Gaussian processes. The relation between the predictions **y**=GP(**x**), **y**'=GP(**x**') is encoded in the covariance function or kernel k(**x**, **x**') of the Gaussian process. For example, we can encode that two distant values **y,y'** of the observations **x**, **x**' will have more different prediction than closer ones. Moreover, we can specify via those covariance functions k(**x**, **x**') and their hyper-parameters **H** the degree of smoothness of the function, the importance of each variable in the prediction *y* or if the function *y*=f(**x**) is stationary or periodic. Critically, we find that just applying the generic squared exponential covariance function works well for our considered problems. We leave for further work to design more complex covariance functions to improve the performance of Gaussian processes in our experiments.

Finally, we would like to briefly describe the basic idea of an additional method based on an ensemble of decision trees called random forest (Breiman, 2001). In particular, decision trees are machine learning methods that are easy to train but lack of strong predictive performance. Random forests are basically combinations of decision trees where every combination focuses on different samples of observations chosen randomly with replacement, targeting different areas of the dataset D={**X**,**y**} and improving drastically the prediction *y* with respect to a single classifier. Finally, the prediction can be done by the majority rule in the case of classification or the empirical predictive distribution of the different classifiers in the forest in the case of regression.

As we have seen, there exists a lot of machine learning models that we summarize in the **M** set. Moreover, each of them have its own set of assumptions that are needed to work. For example, linear regression assumes that data is homocedastic. That is, that the noise *e* of the target variable *y* is constant for all the range of the explanatory variables **x**. Gaussian processes assume that the variable *y* is stationary, which means that the variability of the predictor variable as a function of the explanatory variables **x** is constant in all the range of the explanatory variables **x**. If these assumptions are not validated from the data D={**X**,**y**}, the performance of this algorithms may suffer. Furthermore, machine learning algorithms **M** include a set of hyper-parameters **H**m, where m is a particular machine learning algorithm, whose values **h** need to be set by the user before fixing the training set of observations D_train={**X**_train,**y** _train}. For example, including interactions between the explanatory variables **x** in the case of linear regression, like an additional effects of an extra drink xi if you are a woman xj, or choosing the particular covariance function k(**x**, **x'**) in the case of Gaussian processes. As we do not always know how to fix these hyper-parameters **h** of choosing the particular model m, machine learning practitioners usually test several configurations {H1, ...HN} in a grid (testing a summary of combinations) or in a random fashion. More sophisticated procedures such as Bayesian optimization (Garrido-Merchán, 2020) are used for automatic machine learning tools, that only require to be given a particular dataset D={**X**,**y**} and find the optimum model and set of hyper-parameters Hm in terms of the minimization of the estimation of the generalization error to perform the regression or classification task.

3. Experiments

To add empirical evidence to our claim that we can use machine learning in scenarios related with tourism, we have realized two experiments with real data that are explained in the following two subsections, the first one describing the data that we have experimented with and an the second with the performance of the machine learning models in that data.

3.1 Data

In this paper, we use the EGATUR survey of August, 2021 (EGATUR August, 2021) that collects information on 8294 tourists who visited Spain during the month of August, 2021. The survey registers the following data for each tourist: which country they come from (**country**), number of overnight stays (**night_stays**), the Spanish autonomous community in which they have stayed longer (**ccaa**), the mean of transportation by which they are going to leave Spain (**exit**), the type of accommodation they have mostly used (**accommodation**), the main reason of the trip (**reason**), if they have a tourist package (**package**) and, finally, the total expenditure made during the stay (**expenditure**). Each of these basic tourist data is divided into different types and it is recorded using a numerical code. (see Table 1).

3.2 Results

We have computed two mathematical models: a regression model and a classification model. Our regression model estimates the total (**expenditure**) of a tourist knowing the other variables of the EGATUR survey (**country**, **night_stays**, **ccaa**, **exit, accommodation**, **reason** and **package**). Our classification model classifies the total tourist (**expenditure**) into three levels: low (up to 1500 Euros), medium (from 1500 Euros up to 3000 Euros) and high (more than 3000 Euros), depending on the other variables of the EGATUR survey.

3.2.1 Regression model: Exponential GPR

Our regression model has been obtained from the the Matlab Regression Learner application that performs Machine Learning on a total of twenty four different methods grouped into: neural networks, Gaussian Process Regression (GPR), Ensembles of trees, support vector machines (SVM), regression trees and Linear regression models. To train all these models, we have used hold out validation for a percentage of 25%, which is recommended when the number of data is very large.

We have chosen the exponential Gaussian Process Regression (Exponential GPR) because it has the lowest validation RSME (389,86) and, besides, it also has one of the greatest R-squared validation coefficient (0.85). Other technical characteristics of the Exponential GPR can be consulted in Image 1.

The preditions and the residuals of our method are represented in Graphics 1 and 2, respectively. We can observe a high concentration of values arround the perfect prediction and the zero residual.

Finally, we give an example to visualize the accuracy of our model. We have extracted from the EGATUR survey the information of all the tourists (202) that have visited Pais Vasco (**ccaa**=16). For these tourists, we have computed a graphic (see graphic 3), whose coordinate axes are **expenditure** and **night_stays,** and contains the expenditure during their vacation and the predicted expenditure obtained with Exponential GPR depending on the rest of variables.

3.2.2 Classification model: Ensemble (bagged trees)

Our classification model has been obtained from the Matlab Classification Learner application that performs Machine Learning on a total of thirty one different methods grouped into: neural networks, Gaussian Process Regression (GPR), decission trees, support vector machines (SVM), discriminats, Naive Bayes, support vector machines (SVM), K-Nearest Neighbour KNN, Ensemble and Nucleus methods. To train all these models, we have used again hold out validation for a percentage of 25%.

In this case, the chosen model is Ensemble (bagged trees) because it has the highest accuracy in validation (89.4 %). This validation is obtained from the confusion matrix (see graphic 4), computing the success rate (1854/2073) in the prediction of the three levels of expenditure of a tourist (low=0, medium=1 and high=2). Other technical characteristics of our method can be consulted in Image 2.

In addition to the confusion matrix, the accuracy of our model can be also measured from the area below the ROC curve (see graphic 3). The ensemble method involves an area of almost one (0.97).

Finally, as in the previous model, to visualize the accuracy of the Ensemble method, we have chosen again the 202 Pais Vasco tourists (**ccaa**=16) as an example. The prediction of their level of expenditure is given by the confusion matrix of graphic 6. In this case, the validation is 97,5%.

4. Conclusions and further work

Machine learning has revolutionized a lot of different disciplines such as computer science, astrophysics or biology. This success comes from the power of prediction of the value of one variable as a function of other explanatory variables. To do so, machine learning models are fixed from a dataset of previous observations, discovering the patterns that relate the explanatory variables and the endogeneous variable.

We have studied the state of the art of the application of machine learning in tourism, discovering that although it has been used in some contexts as forecasting the demand of cruises or in hotel recommender systems, its application is not as significant as in other fields.

To emphasize the usefulness of the application of machine learning in tourism, we have performed a couple of experiments with real data, modelling the total expenditure of a tourist and classify it as low, medium or high, based on other explanatory variables such as the number of overnight stays or the origin country. The Exponential GPR minimized RSME out of numerous machine learning methods with R-squared 0,85 and, on the other hand, the ensemble model obtained a validation of 89.4% of cases in the confusion matrix .

As further lines of research, we intend to provide a more technical survey of the application of machine learning in tourism accompanied with a benchmark of tourism problems where different machine learning algorithms may be compared. Our purpose is to provide default machine learning models that can be used for different problems as demand forecasting, tourist profiling, tourism marketing campaigns sentiment analysis or recommender systems. In order to do so, we will explore Bayesian optimization to provide default models and hyper-parameters for each task and investigate how different objectives could be simultaneously optimized with Bayesian optimization such as minimization of the generalization error and confidence in our predictions.

REFERENCES

Bishop, C. M. (2006). Pattern recognition. *Machine learning*, *128*(9).

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

Córdoba, I., Garrido-Merchán, E. C., Hernández-Lobato, D., Bielza, C., & Larranaga, P. (2018, October). Bayesian optimization of the PC algorithm for learning Gaussian Bayesian networks. In *Conference of the Spanish Association for Artificial Intelligence* (pp. 44-54). Springer, Cham.

Cornejo-Bueno, L., Garrido-Merchán, E. C., Hernández-Lobato, D., & Salcedo-Sanz, S. (2018). Bayesian optimization of a hybrid system for robust ocean wave features prediction. *Neurocomputing*, *275*, 818-828.

EGATUR August (2021) Tourist expenditure survey of August, INE (National Institute of Statistics of Spain), link:

https://www.ine.es/dyngs/INEbase/es/operacion.htmc=Estadistica_C&cid=1254736177002&menu=resultados&idp=1254735576863#!tabs-1254736195390

Garrido-Merchán, E. C., & Hernández-Lobato, D. (2020). Dealing with categorical and integer-valued variables in bayesian optimization with gaussian processes. *Neurocomputing*, *380*, 20-35.

Kirilenko, A. P., Stepchenkova, S. O., Kim, H., & Li, X. (2018). Automated sentiment analysis in tourism: Comparison of approaches. *Journal of Travel Research*, *57*(8), 1012-1025.

Le, T. H., Arcodia, C., Novais, M. A., & Kralj, A. (2021). Proposing a systematic approach for integrating traditional research methods into machine learning in text analytics in tourism and hospitality. *Current Issues in Tourism*, *24*(12), 1640-1655.

Murphy, K. P. (2012). *Machine learning: a probabilistic perspective*. MIT press.

Nilashi, M., Bagherifard, K., Rahmani, M., & Rafe, V. (2017). A recommender system for tourism industry using cluster ensemble and prediction machine learning techniques. *Computers & industrial engineering*, *109*, 357-368.

Rasmussen, C. E. (2003, February). Gaussian processes in machine learning. In *Summer school on machine learning* (pp. 63-71). Springer, Berlin, Heidelberg.

Sun, S., Wei, Y., Tsui, K. L., & Wang, S. (2019). Forecasting tourist arrivals with machine learning and internet search index. *Tourism Management*, *70*, 1-10.

VanderPlas, J., Connolly, A. J., Ivezić, Ž., & Gray, A. (2012, October). Introduction to astroML: Machine learning for astrophysics. In *2012 conference on intelligent data understanding* (pp. 47-54). IEEE.

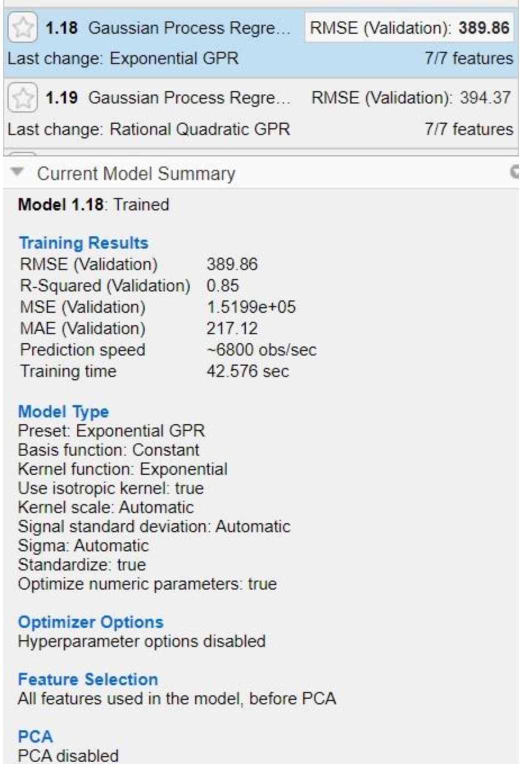
Xie, G., Qian, Y., & Wang, S. (2021). Forecasting Chinese cruise tourism demand with big data: An optimized machine learning approach. *Tourism Management*, *82*, 104208.

Tables, graphics and images

Table 1 Description of variables (Section 3.1)

Column	Meaning	Classes
exit	Way of exit of Spain	1: road 2: airport, 3: port, 4: train
country	Country of habitual residence	01: Germany. 02: Belgium. 03: France. 04: Ireland. 05: Italy. 06: Netherlands. 07: Portugal. 08: United Kingdom. 09: Switzerland. 10: Russia. 11: Nordic countries (Denmark, Finland, Norway, Sweden). 12: Rest of Europe. 13: USA. 14: Rest of America. 15: Rest of the world
ccaa	Autonomous community of Spain	01: Andalucía. 02: Aragón. 03: Principado de Asturias. 04: Illes Balears. 05: Canarias. 06: Cantabria. 07: Castilla y León. 08: Castilla-La Mancha. 09: Cataluña. 10: Comunitat Valenciana. 11: Extremadura. 12: Galicia. 13: Comunidad de Madrid. 14: Región de Murcia. 15: Comunidad Foral de Navarra. 16: País Vasco. 17: La Rioja. 18: Ceuta. 19: Melilla
night_stays	Overnight stays	
accommodation	Main accommodation	1: Hotels, 2: Rest of the market, 3: Non-market accommodation
reason	Main reason for the trip	1: Vacation, 2: Business, 3: Rest
package	Tourist Package	1: Yes, 0: No
expenditure	Total expenditure per person	

Image 1 Technical characteristics of Exponential GPR (Section 3.2.1)



1.18 Gaussian Process Regre... RMSE (Validation): **389.86**
Last change: Exponential GPR 7/7 features

1.19 Gaussian Process Regre... RMSE (Validation): 394.37
Last change: Rational Quadratic GPR 7/7 features

Current Model Summary

Model 1.18: Trained

Training Results

```
RMSE (Validation)        389.86
R-Squared (Validation)   0.85
MSE (Validation)         1.5199e+05
MAE (Validation)         217.12
Prediction speed         ~6800 obs/sec
Training time            42.576 sec
```

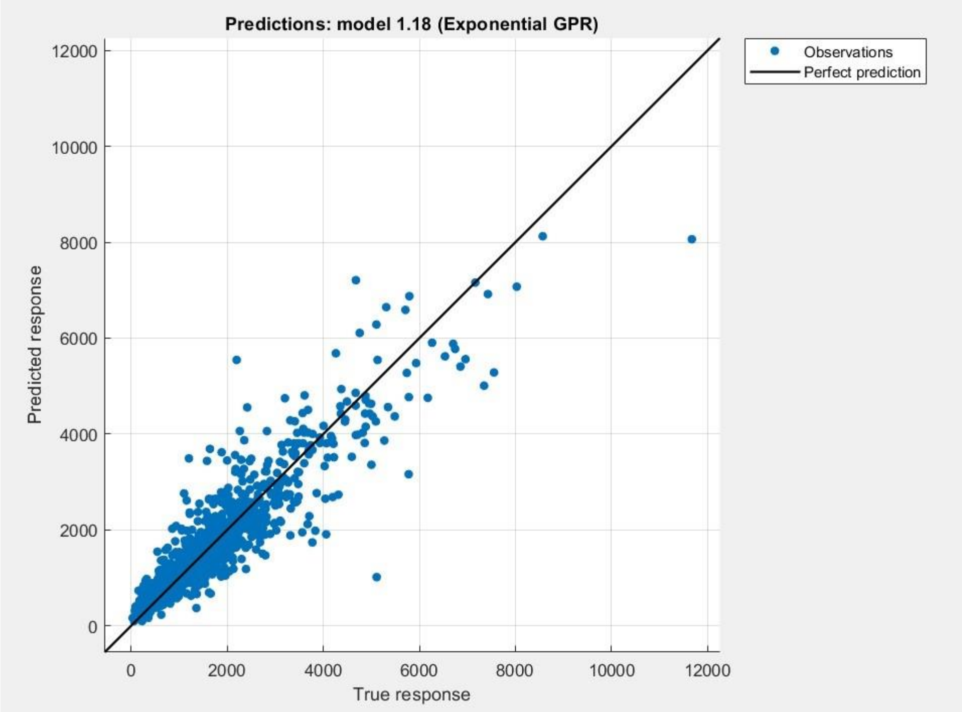
Model Type
Preset: Exponential GPR
Basis function: Constant
Kernel function: Exponential
Use isotropic kernel: true
Kernel scale: Automatic
Signal standard deviation: Automatic
Sigma: Automatic
Standardize: true
Optimize numeric parameters: true

Optimizer Options
Hyperparameter options disabled

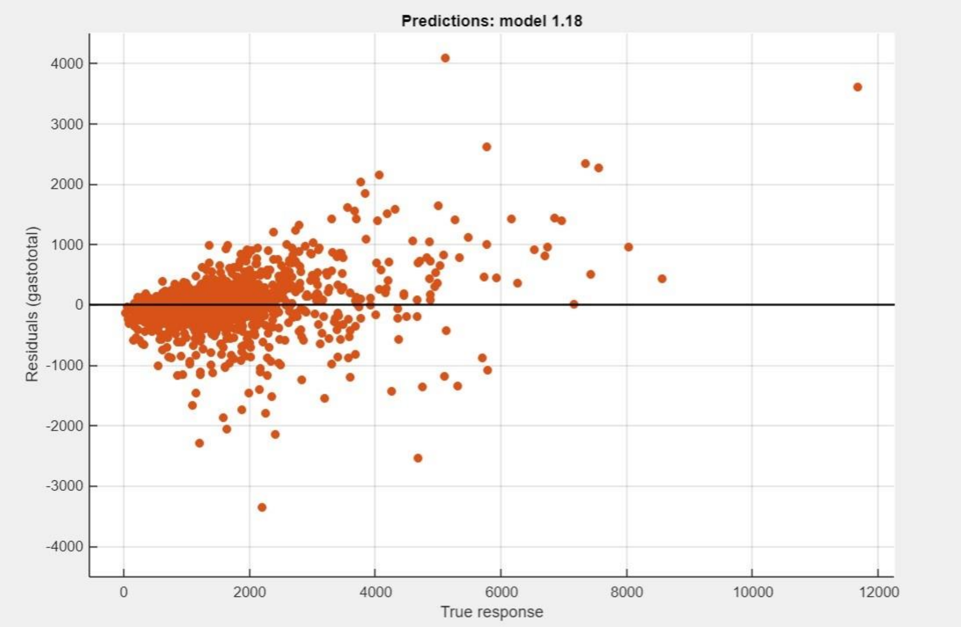
Feature Selection
All features used in the model, before PCA

PCA
PCA disabled

Graphic 1 Predictions of Exponential GPR (Section 3.2.1)



Graphic 2 Residuals of Exponential GPR (Section 3.2.1)



Graphic 3 Relation between **e**xpenditure and overnights stays of turism in País Vasco (real data and predictions of exponencial GPR) (Section 3.2.1)

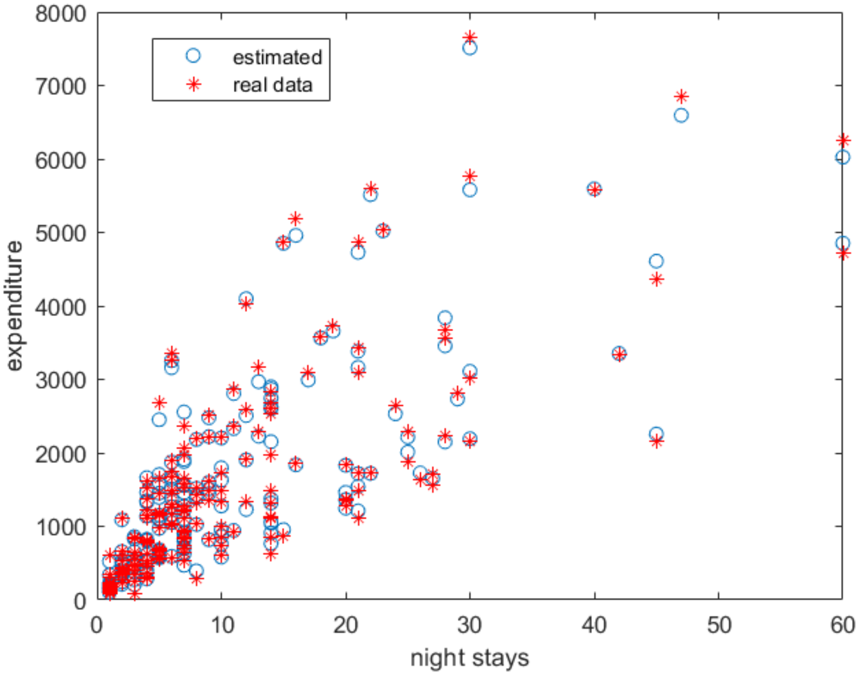
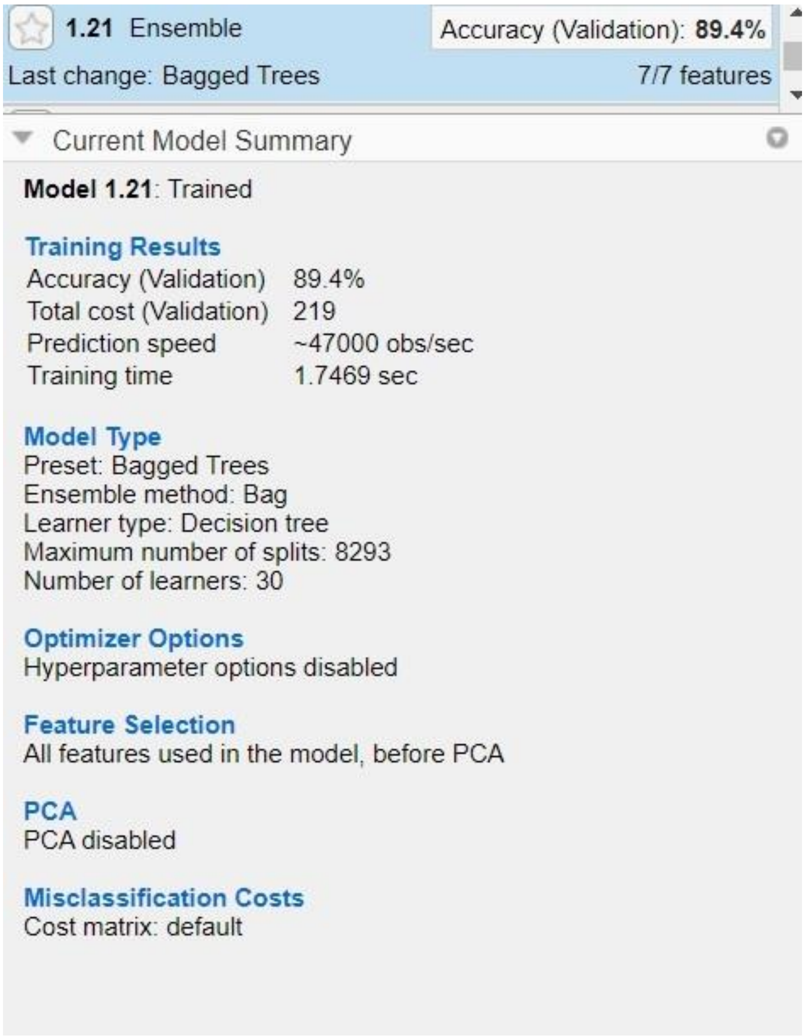


Image 2 Technical characteristics of Ensemble (bagged trees) (Section 3.2.2)



1.21 Ensemble

Accuracy (Validation): **89.4%**

Last change: Bagged Trees

7/7 features

Current Model Summary

Model 1.21: Trained

Training Results

Accuracy (Validation)	89.4%
Total cost (Validation)	219
Prediction speed	~47000 obs/sec
Training time	1.7469 sec

Model Type

Preset: Bagged Trees
Ensemble method: Bag
Learner type: Decision tree
Maximum number of splits: 8293
Number of learners: 30

Optimizer Options

Hyperparameter options disabled

Feature Selection

All features used in the model, before PCA

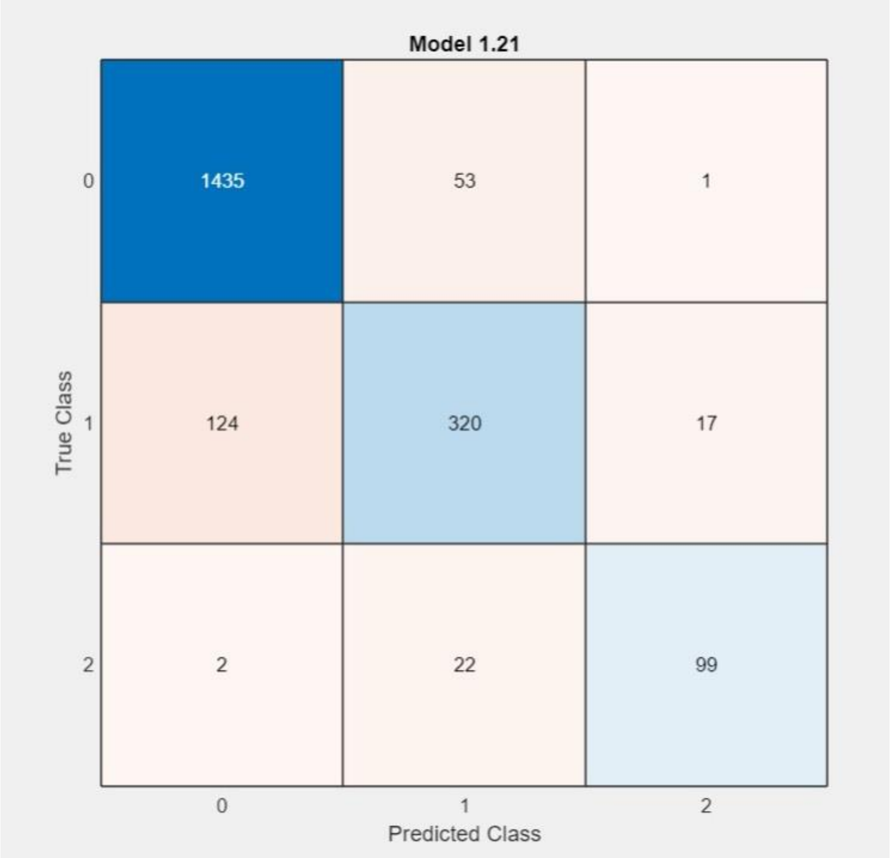
PCA

PCA disabled

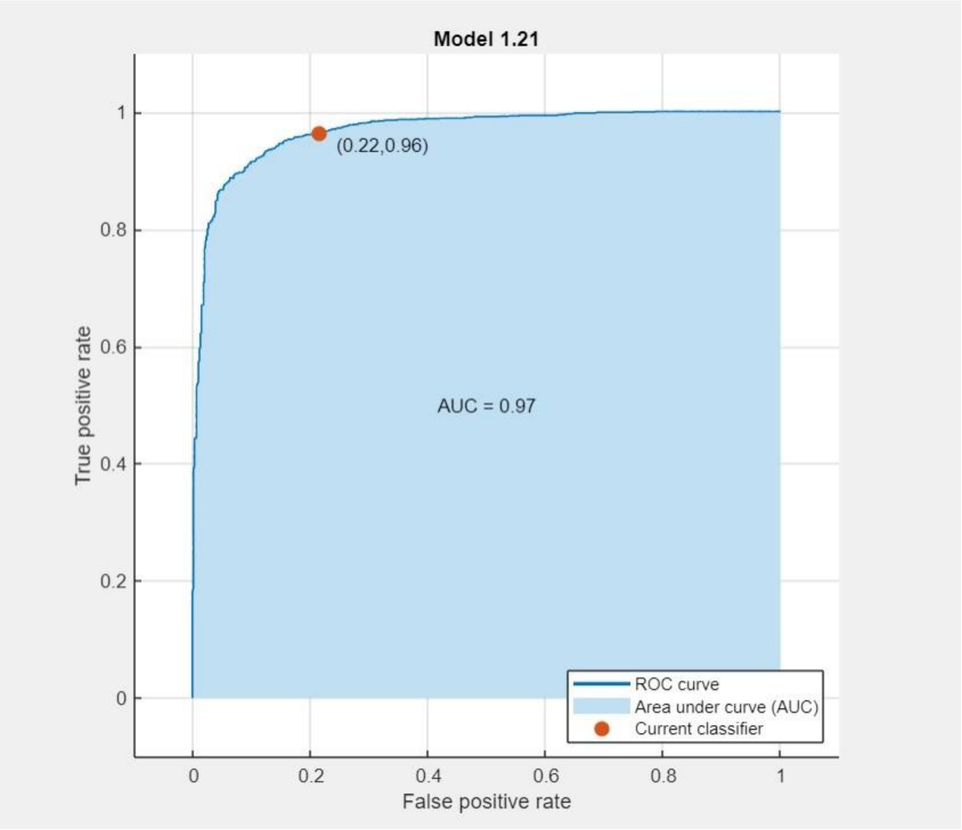
Misclassification Costs

Cost matrix: default

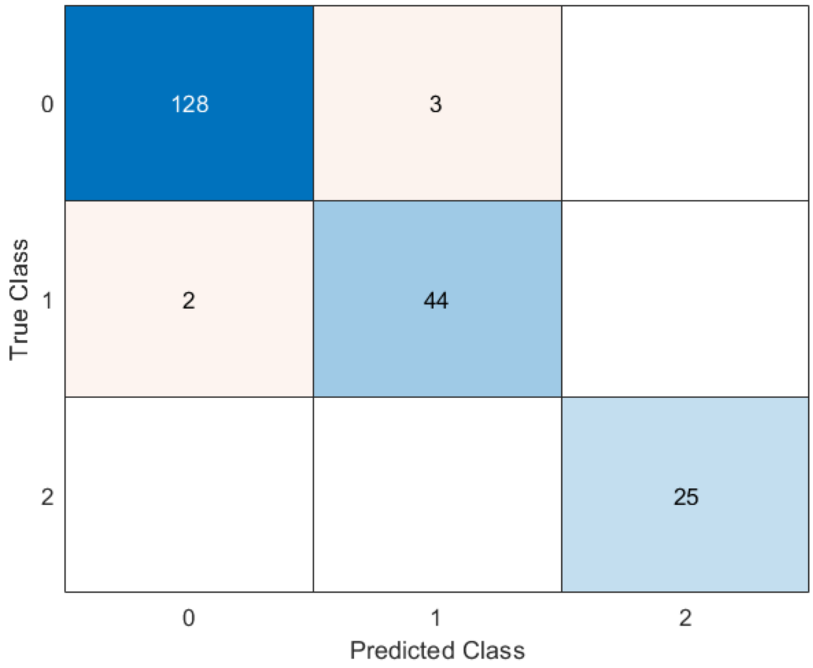
Graphic 4 Confusion matrix of Ensemble (bagged trees) (Section 3.2.2)



Graphic 5 Roc curve of Ensemble (bagged trees) (Section 3.2.2)



Graphic 4 Confusion matrix of Ensemble (bagged trees) for País Vasco tourists (Section 3.2.2)



True Class	0	1	2
0	128	3	
1	2	44	
2			25