Predicting the participation in kuntorastit events using machine learning

1. Introduction

This project applies machine learning (ML) to predict the number of participants in kuntorastit events. Kuntorastit are recreational orienteering events organized around Finland every week from April to October (see e.g. https://www.rastilippu.fi/). They are hosted by different orienteering clubs on a volunteer basis and usually gather around 100-500 participants per event. The number of participants is important information to be used in the planning and organizing of the events. For example, the participation affects the demand for maps, required number of volunteers, and parking arrangements. The aim of this project is to find a model that can better predict the participation in future events than the "educated guess" method applied now. The report is organized as follows: Section 2 formulates the task as a ML problem and present the data. Section 3 discusses the methods applied in solving the ML problem. The fourth section presents the results obtained with different ML models. Finally, conclusion summarizes the main findings.

2. Problem formulation

This section formulates the task as a ML problem and presents the data used in the project. To find a model to predict the number of participants in future events I first train different ML models using data from past events. The data points are individual kuntorastit events such as Espoorastit 2.7.2019 in Leppävaara, Espoo. I use data from hundreds of events organized by Espoon Akilles orienteering club between 2009-2019. The labels of the data points are the total number of participants in each event (see Figure 1). This information together with some of the feature variables is provided by Espoon Akilles.

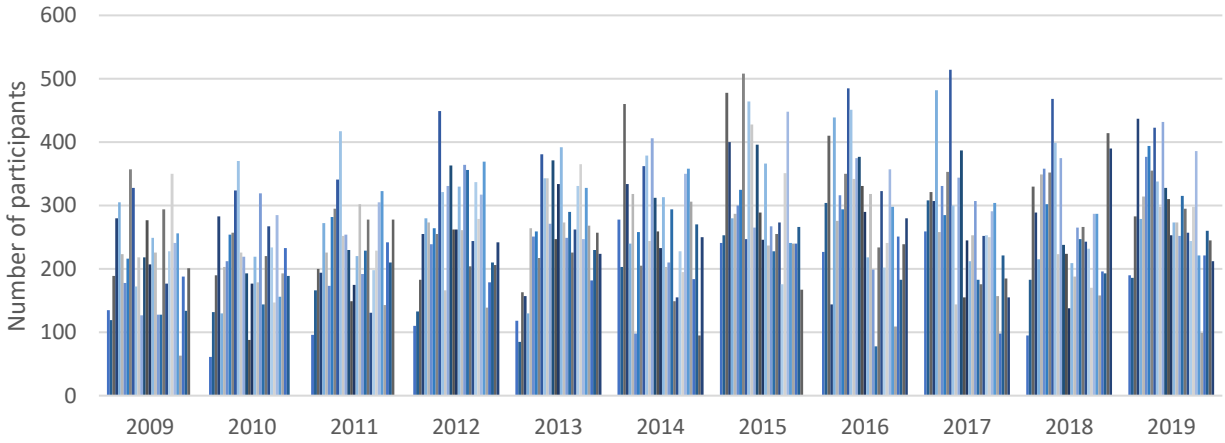

Figure 1 Participation in Espoorastit events 2009-2019
This figure shows the number of participants in each Espoorastit event (columns) from 2009 to 2019.

Table 1 lists all data sets and variables used in this project. The features of the data points include details on event time, location, and weather which are expected to affect the participation. I use the year and month of the event and weeks to next Jukola relay as time features as there seems to be some seasonality in participation. For location features I use expert evaluations of the event locations' optimality and difficulty level. Some locations seem to constantly attract more participants which is hopefully captured by these variables. Finally, as orienteering is an outside sport, I expect precipitation and temperature to affect the participation. The predictions of different ML models are based on these feature variables.

Table 1 Datasets used in the project

Data	Description	Type	Source
Participation (label)	Number of participants in each event	Numeric	Espoon Akilles : https://www.espoonakilles.fi/espoorastit
Year	Year the event is organized in	Numeric	
Month	Month the event is organized in	Numeric	
Weeks to Jukola	Number of weeks to next Jukola relay	Numeric	Jukola: https://jukola.com/en/
Optimality	Event location's suitability/popularity (-1 (bad), 0 (ok), or 1 (good))	Categorical	Expert option
Difficulty level	Event location's difficulty level (-1 (easy), 0 (normal), or 1 (hard))	Categorical	Expert option
Precipitation	Precipitation (mm) on the day of the event	Numeric	Finnish meteorological institute: https://en.ilmatieteenlaitos.fi/download-observations
Temperature	Maximum temperature (°C) on the day of event	Numeric	

3. Methods

This section explains the ML models and methods used in the study. Beginning with the models, I test two types of ML models to predict the participation: linear regression and k-nearest neighbors regression. As the participation (label) is a numeric and continuous quantity, regression-based models are a suitable category of ML methods and often used in practice (Jung, 2021). The models are explained below.

Linear regression

- uses linear hypothesis maps h(**x**) and fits a multivariate linear model with coefficients w = w1, w2, …, wn for features **x** plus an intercept term w0
- loss function minimizes the sum of squared errors to find the optimal coefficients for the model
- scikit-learn reference: sklearn.linear_model.LinearRegression

K-nearest neighbors regression

- predicts the label by calculating the average label of k-nearest neighbors (i.e. most similar events)
- parameters: number of neighbors (k), weights (uniform or relative), algorithm to find neighbors
- no loss function, train and validation errors are calculated using the k-neighbors from training set
- scikit-learn reference: sklearn.neighbors.KNeighborsRegressor

In addition to using two different types of models, I use different specifications of both by varying the number of features used in the training. For example, I train and validate models using only time, time + location, or time + location + weather features. This gives me further insights on the performance of the models and also the effect of different feature variables on the participation. For the k-nearest neighbors regression, I also test models with different number of neighbors and alternative weighing schemes.

To train and evaluate the models I split the data into two datasets: events in 2009-2018 and events in 2019 which consists of 305 and 31 datapoints, respectively. I use the 2009-2018 data to train and validate different ML models and to choose the best model which achieves the lowest validation error measured by squared error loss (Jung, 2021). Because the features variables are on different scales, I standardize the dataset using scikit-learn's StandardScaler tool. Furthermore, as the sample size is fairly limited, I use k-fold cross-validation (k-fold CV) with k = 5 to train and validate the models. K-fold CV produces an average validation score from k train/test sets for each model. Finally, I asses the prediction quality of the best performing ML model using the 2019 events and comparing the predicted participation with the actual participation.

4. Results

This section presents the results of the study and the best model for predicting the participation in kuntorastit events. Beginning with the performance of different ML models, Table 2 shows the training and validation errors of linear regression and k-nearest neighbors regression models with different sets of features. Linear regression models tend to have higher training errors than k-nearest neighbors models, which is due to their methodological differences. However, linear regression models have lower validation errors meaning that they perform better that k-nearest neighbors regression models (Jung, 2021).

Looking at the training and validation errors obtained with different sets of features (combinations of time, location, and weather features), we see that the timing of the event seems to be the best predictor for participation (lowest training and validation error) followed by location and weather. For the k-nearest neighbors regression, varying the number of neighbors and weighing scheme has some effect for the results. A smaller value of k leads to larger training and validation errors whereas 'distance' weighing scheme achieves somewhat lower errors (results untabulated). However, linear regression models are still better.

Table 2 Training and validation errors of different models
This table shows the results from k-fold cross validation for linear regression and k-nearest neighbors regression models with different sets of features. The model with the lowest validation error is highlighted.

Model	Features	Training error	Validation error
Linear regression (incl. intercept)	time[1]	5,006	6,434
	location[2]	6,270	6,847
	weather[3]	6,850	7,466
	time + location	4,307	5,794
	location + weather	5,915	6,533
	time + weather	4,856	6,373
	time + location + weather	4,192	5,763
K-nearest neighbors regression (k = 5, weights = uniform, algorithm = auto)	time	3,489	5,852
	location	7,528	8,198
	weather	5,284	8,681
	time + location	3,643	6,203
	location + weather	4,866	7,825
	time + weather	3,841	6,091
	time + location + weather	3,616	5,945

1) time refers to month and year of the evet, and weeks to Jukola; month is used as a dummy variable in the linear regression (e.g. a separate variable (1/0) for every month) and as a continuous variable (e.g. month 4-10) in the k-nearest neighbors regression
2) location includes evaluations of the optimality and difficulty level of the event location
3) weather refers to precipitation and maximum temperature on the day of the event

Based on the comparison of validation errors, the best model for predicting the participation is a linear regression model with time, location, and weather features. This model achieves an average validation error of 5,763 for k-fold CV which is lower than that of other models'. The best k-nearest neighbor model is the one with only time features and this model's k-fold CV error is 5,852.

To assess the performance of the best model, I test it by predicting the 2019 event participation and comparing the results with the actual, realized participation. Figure 2 plots the predictions for the 31 events in 2019. The model performs quite well and achieves a testing error of 3,466. The average error per event is only 3.8 but the largest absolute errors are over 100. Especially the predictions for events in the spring and early summer (events 4-13) and in location Leppävaara (events 4, 14 and 25) are off by a lot.

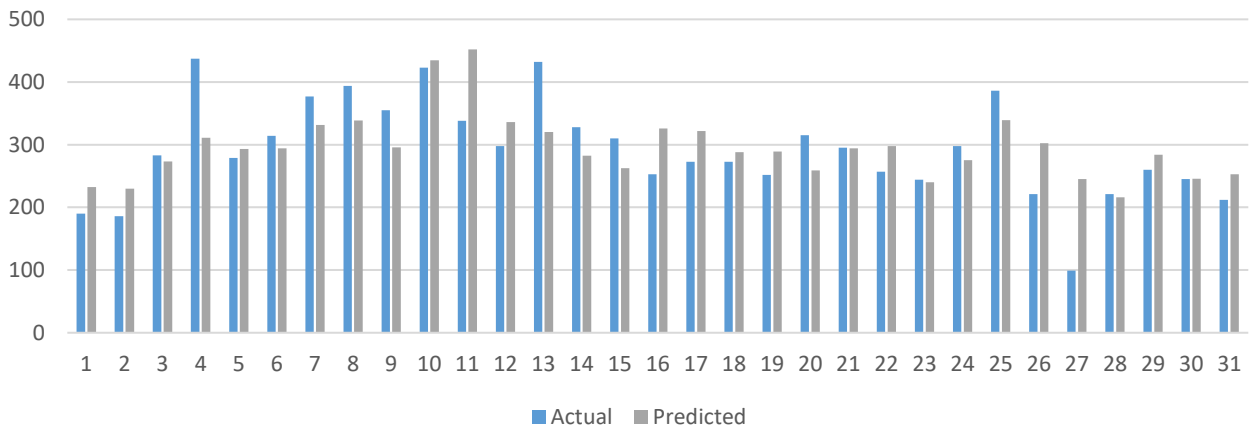

Figure 2 Actual and predicted participation in 2019 events
This figure plots the actual and predicted participation in 31 Espoorastit events in 2019. The predicted participation is estimated with a linear regression model with location, time, and weather features.

5. Conclusion

In this project, I have applied ML to predict the participation in kuntorastit events. My data consisted of Espoorastit events from 2009-2019 where the labels are the total number of participants in each event and the features consist of event time, location, and weather statistics. I tested two types of ML models, linear regression and k-nearest neighbors regression, with k-fold CV and compared the training and validation errors of the models. The best performing model is a linear regression model with time, location, and weather features which achieved the lowest k-fold CV error of 5,763 on the training set. Finally, I tested the best model on the test set where it achieved a validation error of 3,244.

Going forward, the performance of the ML model could be improved by increasing the sample size and including more features to the data. The current project was done using only Espoorastit events from 2009-2019 but there are dozens of other kuntorastit organizers with data from thousands of events. The features were now limited to time, location, and weather statistics but including for example event organizer, more location characteristics (latitude/longitude coordinates, availability of public transport, etc.) and more accurate weather data could improve the ML model. However, the performance of the current, best model is already relatively good and certainly better than the "educated guess" method applied in the estimation previously.

Bibliography/References

Espoon Akilles (2021). Espoorastit. https://www.espoonakilles.fi/espoorastit

Jung, A. (2021). Machine Learning: The Basics. Available at mlbook.cs.aalto.fi.

Scikit-learn (2021). Machine Learning in Python. Available at https://scikit-learn.org/stable/index.html