Predicting number of visitors at Korkeasaari with regression methods

Introduction

Great planning can be very hard but sometimes very important. The success of a plan may be important on a personal level but also from a business perspective. Maybe you would like to attend an event, but only in case the event is really popular or maybe you would rather go visit some place when it's very quiet so you can enjoy everything at your own peace. You may want to understand how busy your business is going to be, so you know how many employees need to have working during a certain period. For these examples predicting how many visitors does some place have based on some features would be very important.

In this project we are trying to offer a solution to this problem of predicting the number of visitors based on some features. We have picked Korkeasaari Zoo as the example case as we were able to find the kind of data that we think is required for the problem on Korkeasaari Zoo. In this example case we are going to predict the number of visitors at Korkeasaari with machine learning regression methods.

Our project report consists of seven chapters, including the introduction chapter. In the second chapter we are going to go over the problem formulation. In the third chapter we talk about the methods we are using. The fourth chapter focuses on showing the results of our project. In the fifth chapter we discuss the conclusion of our project. The references used in the report can be found from the sixth chapter and appendices from the seventh chapter.

Problem formulation

The goal is to predict the number of visitors at Korkeasaari using machine learning and more specifically, regression methods. The task is a supervised learning problem. Each data point represents a day in Korkeasaari, listing the following details: weekday, month, air temperature, rain status and number of visitors. For features we are using month, weekday, air temperature and rain status. The label is the number of visitors. Month and weekday are categorical variables. Air temperature is a continuous numerical variable, which represents the air temperature of the day in celsius. Rain status is a binary variable that represents whether it rained or not during that day. The number of visitors is a continuous numerical variable that represents the number of visitors during the entire day. The dataset has been formed using statistics on the number of visitors in Korkeasaari (Helsinki region infoshare, 2023) and statistics on the weather data (Ilmatieteen laitos, 2023). More specifically, the daily visitor statistics of Korkeasaari from 2013-2022 and daily weather statistics collected in Helsinki, Kaisaniemi weather station from 2013-2022 were used.

Methods

The dataset is created by combining two datasets with each other and then filtering out anomalies. The code used to do this can be found at the Appendices. At first we extracted visitor data from multiple Korkeasaari visitor statistics. Then we combined this with the Ilmatieteen laitos weather statistics. As the visitor statistics may have periods that the Korkeasaari might not have had as many visitors as usual due to some anomaly (COVID-19, Renovations, etc.), we decided to locate and remove these anomalies from the dataset. We used monthly visitor counts to calculate monthly averages and located month/year pairs that had below 1/10th of the average visitors for that month. Then we filtered the dataset by removing entries with month/year pairs that matched the pairs we had located. Then we dropped columns that stored unnecessary information. From this process we received a dataset that has 3441 data points that follow the formation described in section Problem formulation.

The features were decided by thinking of what factors affect the decision to go to Korkeasaari and then analyzing those features by thinking which of these can be obtained easily. We thought that month and weekday are factors that affect whether people visit Korkeasaari and they are easily obtainable. We also thought that the weather affects the decision, so air temperature and rain status was decided to be used as features, as they are also easy to measure or obtain.

Linear regression was decided as the first machine learning method to be used. Linear regression establishes a linear relationship between the features and labels. It was chosen because it is a simple but still very effective machine learning method for regression problems. We also think that the relationship between features and labels can be represented as linear, especially with one-hot encoding for categorical variables.

Multi-layer perceptron was decided as the second machine learning method to be used. It can capture complex non-linear relationships between the features and labels using multiple layers of interconnected neurons. The reason why it was chosen as the second method was exactly because it can capture complex non-linear relationships between the features and labels. By using both linear regression and multi-layer perceptron we should end up with good results no matter whether the relationship between the features and labels is linear or non-linear.

Mean squared error was decided as the loss function to be used with linear regression and multi-layer perceptron. The mean squared error measures the average squared difference between the predicted label and the actual label for the given features. It was chosen because it is sensitive to errors due to the squaring of the error (makes large errors even larger) and because of its efficiency as it is quick to calculate. The mean squared error works well with both linear regression and multi-layer perceptron and is usually the preferred loss function that is used with them.

For the model validation the dataset is first shuffled to ensure the data points are in random order to minimize the effect of possible biases. Then the dataset is split into training/validation and test sets by 85%/15% ratio. For the testing 15% is chosen because it

is a regularly used value in machine learning and it still leaves most of the data to be used for training and validation. Then for the 85% training/validation split k-fold validation is used with 6 splits. The k-fold validation is used because it improves the model's estimates compared to a single split being used for training and validation. The reason for choosing 6 splits is that when used with the 85% testing/validation split it offers training and validation splits for each fold that are close to a single-split training and validation splits of ratio 70%/15%.

Results

The linear regression model and the multi-layer perceptron model were trained with the split dataset. The splitting of the dataset and validation during the training was done as described in the Methods chapter. For the testing, the testing split constructed in the Methods chapter was used. The test split consists of data points that have neither been used during the training or validation. The code used to do the training can be found at the Appendices. The trained linear regression model had a training error of 1142721, a validation error of 1159455 and a test error of 1037030. The trained multi-layer perceptron had a training error of 952908, a validation error of 1037753 and a test error of 984354.

The linear regression model had larger training and validation error compared to the multi-layer perceptron model. Also the test error of the linear regression model is larger than the test error of the multi-layer perceptron model. This means the multi-layer perceptron model performs better when tasked to predict labels for previously unseen data. For this reason we decided that the final chosen method is the multi-layer perceptron with the test error of 984354.

	Training error	**Validation Error**	**Test error**
Linear Regression	1142721	1159455	1037030
Multi-layer perceptron	952908	1037753	984354

Conclusions

In this project we tried to construct a solution to the problem of predicting the number of visitors based on some features. We picked Korkeasaari Zoo as the example case. We first constructed a dataset for this problem with 3441 data points. We then decided on the machine learning methods that we would try out. The chosen methods were linear regression and multi-layer perceptron. We decided to use mean squared error as the loss function for both of the models. We then formulated the dataset splits used in this project. The training/validation split consisted of 85% of the data points and test split of 15% of the data points. For validation k-fold validation was used with 6 splits. We then trained, validated and tested both of the chosen models. After comparing the results of both trained models, we picked multi-layer perceptron as the final chosen method, as it performed better on unseen data based on the lower test error it received.

For the linear regression model the fact that training and validation errors are very close to each other suggests that the model is underfitting. This can be either because the model is not complex enough to capture the patterns in the data set or the data set requires additional features. On the other hand the fact that the validation error was clearly larger than the training error for the multi-layer perceptron suggests that the model is overfitting. This can be either because the model is too complex or because the data set requires additional features.

The results we received in this project don't seem to be optimal. The training, validation and test errors were quite high. The magnitude of the errors seems larger than it actually is because of using mean squared error. Still there is much room for improvement. Our hypothesis is that the used features were not good enough for more precise predictions. Either better features or more features would have been required. With the current features the predictions could have possibly been better if instead of predicting the number of visitors we would have divided the number of visitors into categories (very quiet, quiet, moderately busy, very busy) and predicted the category classification instead. This could be one future direction on how to further improve our machine learning method.

References

Helsinki region infoshare. (2023, January 11). *Korkeasaaren kävijämäärät*. Helsinki Region Infoshare. Retrieved September 22, 2023, from https://hri.fi/data/fi/dataset/korkeasaaren-kavijamaarat

Ilmatieteen laitos. (2023). *Avoin data*. Ilmatieteen laitos. Retrieved September 22, 2023, from https://www.ilmatieteenlaitos.fi/avoin-data

Appendices

Project code and resources are located at Github ML2023_Project.