trivago

Machine Learning Challenge

Overview

Objective

Build a model that predicts the number of hits per session based on a database of user sessions of Trivago

Methodology used

Cross Industry Standard Process for Data Mining (CRISP-DM)

- Business Understanding
- Data Understanding
- Data Preparation
- Modeling
- Evaluation
- Deployment (not applied on this case)

Results

After testing some models, the finalist was a Gradient Boosting with the best Mean Squared Error

Project highlights

trivago

Business Understanding

Research about trivago and deep understanding about the project objective

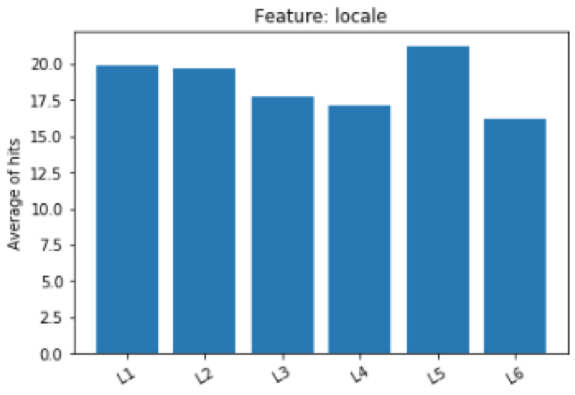
Data Understanding

I read the provided metadata and did an exploratory analysis:

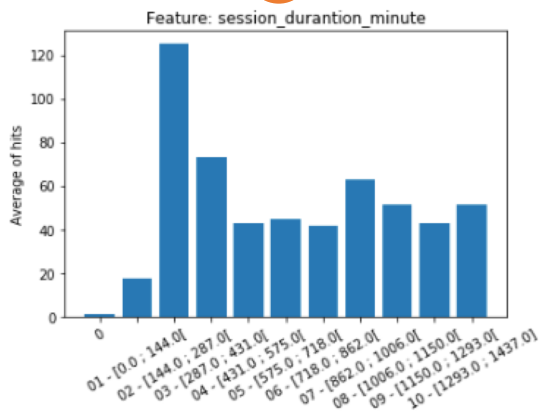
- feature format (string or numeric)
- feature type (discrete or continuous)
- feature distribution

Two examples of feature distribution (one discrete and one continuous):

1



2



1 category 'L5' has an average of hits a little bit higher than the rest of the categories

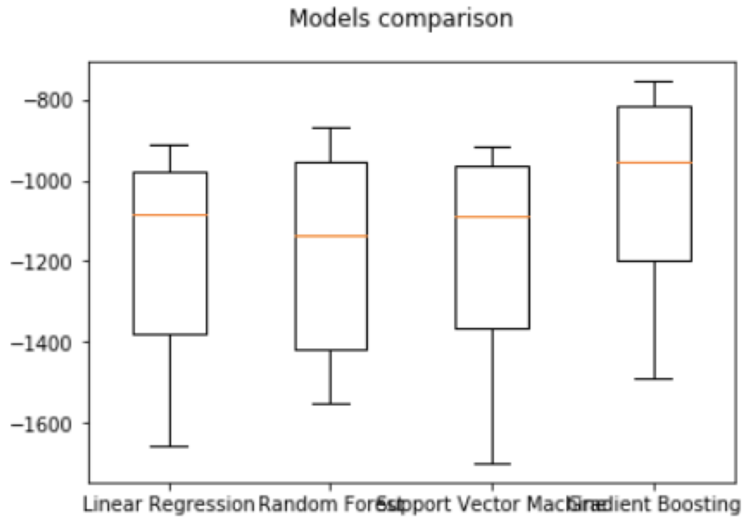
2 band '01' that contemplates between 0 and 144 minutes has a high average of hits

Data Preparation

- Creation of new features, example: a feature that counts how many paths id were visited, from the original feature 'path_id_set'
- Handle data so it's suitable for applying any model from the python package used and split the data in training and test datasets

Modeling

Training of four different models using k-fold cross-validation and selected the one with the lowest mean squared error (gradient boosting). Tuned the parameters of **gradient boosting** and applied to the test dataset (**with MSE of 930 (training) and 991 (test)**)



Evaluation

Scored the result dataset using the model created in the modeling step

How to enhance?

feature engineering

normalization of features

other distribution analysis

other models

* Python Cross-validation uses negative squared error, that's why the best is the highest

About me

Bio:

I'm from Brazil, I've 5+ years of experience in data science, bachelor in Mathematics and master in Statistics.

I speak English, Spanish (lived for a year in Spain), Portuguese (native) and starting to learn (up to 12 months now) German using Rosetta Stone app.

Info:

lucas.bruscato@gmail.com
linkedin.com/in/lucasbruscato
github.com/lucasbruscato
+55 11 985071210

Plus:

You can find the project and this presentation on https://github.com/lucasbruscato/trivago