Trivago challenge

The number of hits prediction

By: Phuong Pham

Content

- Problem
- Data Visualization
- Data mining pipeline
  - Data processing
  - Feature engineering
  - Training phase
  - Results
- Further improvement

Problem

- Task: Predicting the number of hits (regression problem)
- 10 initial input variables
- ~ 1 million observations
- Evaluation matrix: root mean square error

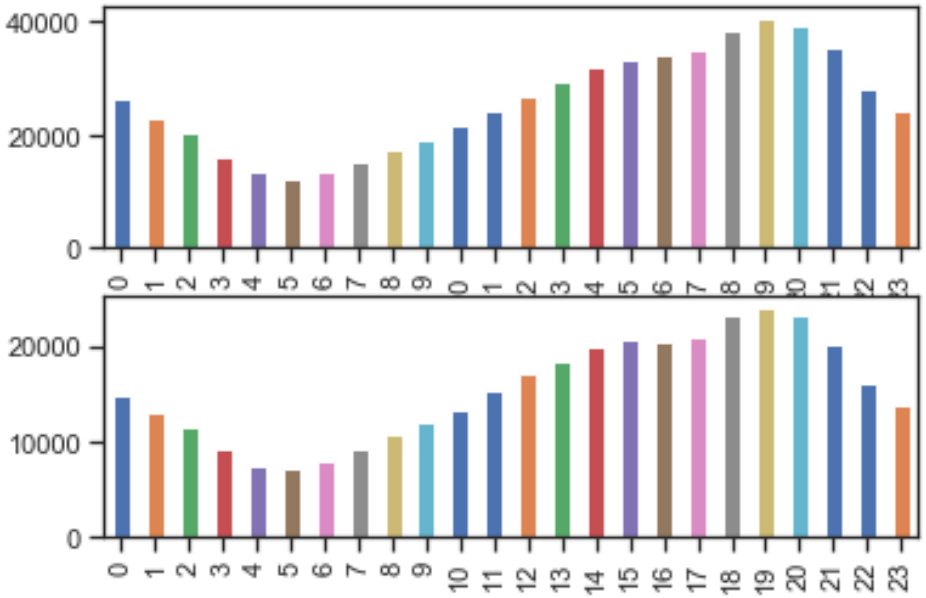
Data visualization

- varialbe “hits” has missing values=> seperate data into 2 subsets base on missing hits”

  => compare characteristics of 2 datasets

- Missing hit dataset: ~370.000 data points
- Available values hit dataset: ~620.000 data points

- Compare the click on hours of day between 2 datasets

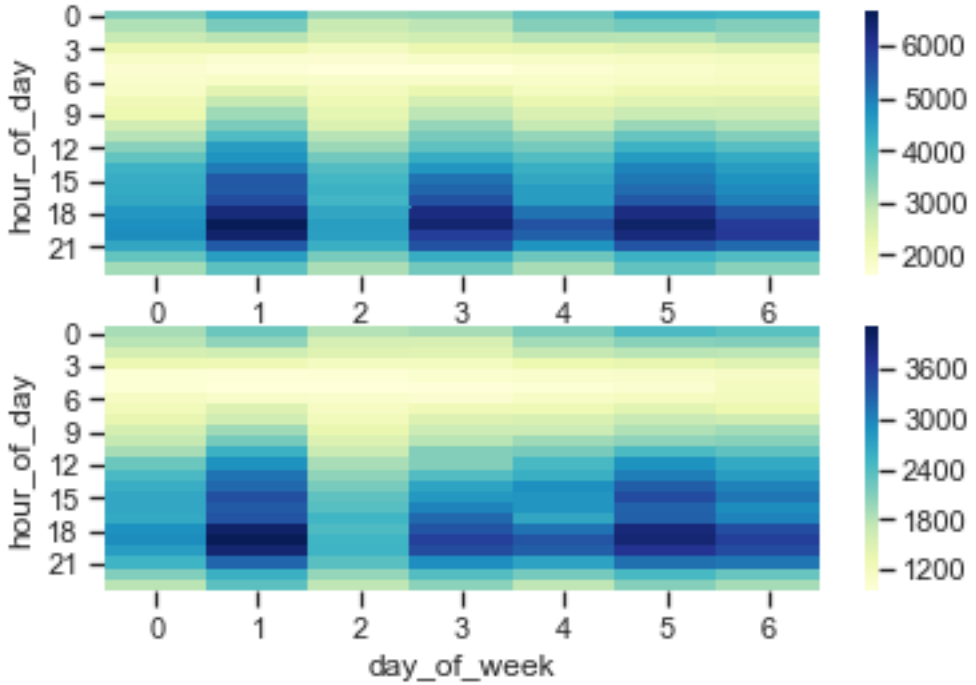


Available values hit dataset

Missing values hit dataset

=> from 17-21 o'clock the number of hits are biggest

- Compare the number of hits based on hour of day and day of week btw 2 datasets:

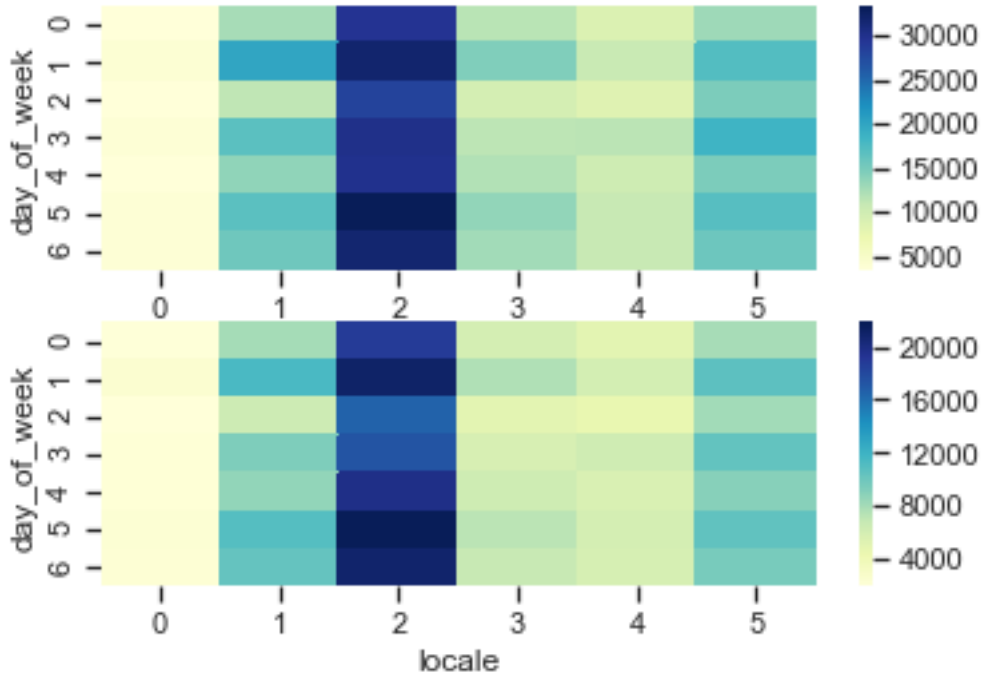


Available values hit dataset

Missing values hit dataset

=> Monday, Wednesday and Friday have highest number of clicks

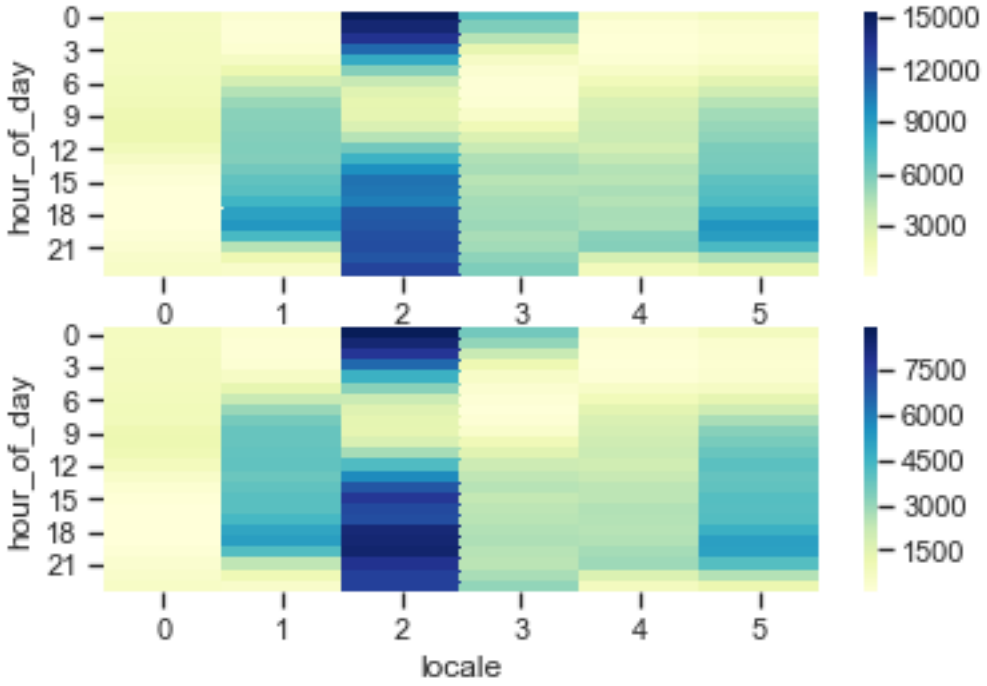
- Compare the number of hits based on locale and day of weeks between 2 datasets:



Available values hit dataset

Missing values hit dataset

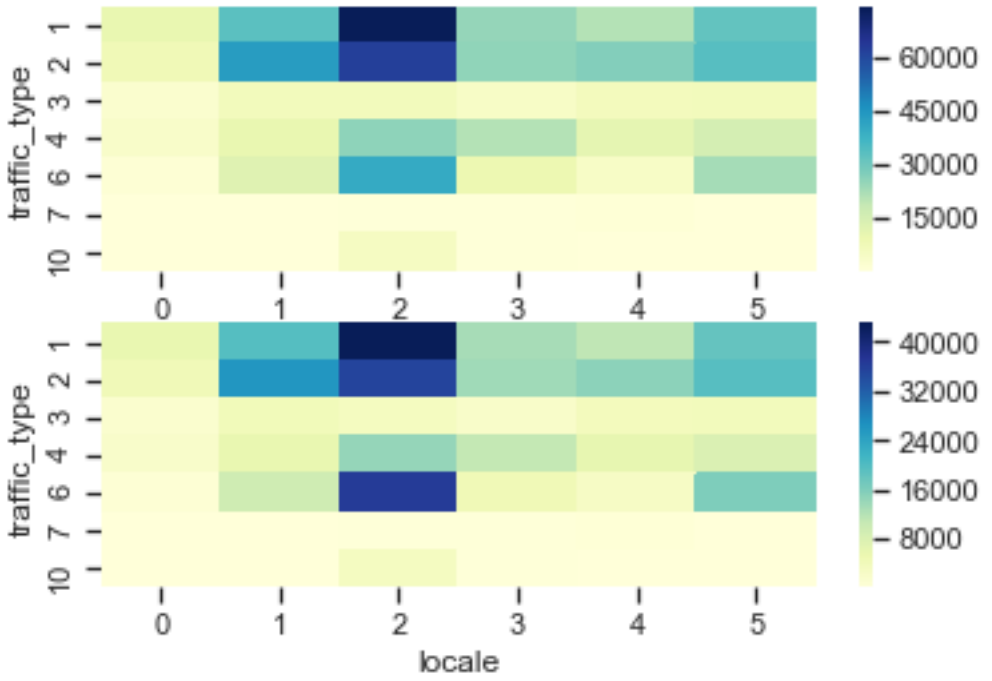
- Compare the number of hits based on locale and hour of day between 2 datasets:



Available values hit dataset

Missing values hit dataset

- Compare the number of hits based on locale and traffic type between 2 datasets:

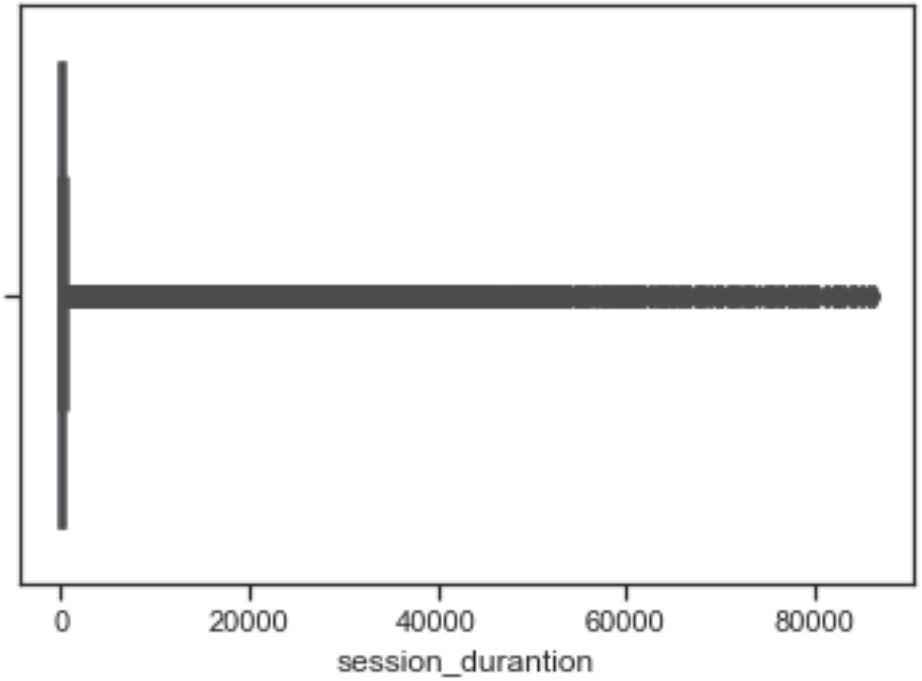


Available values hit dataset

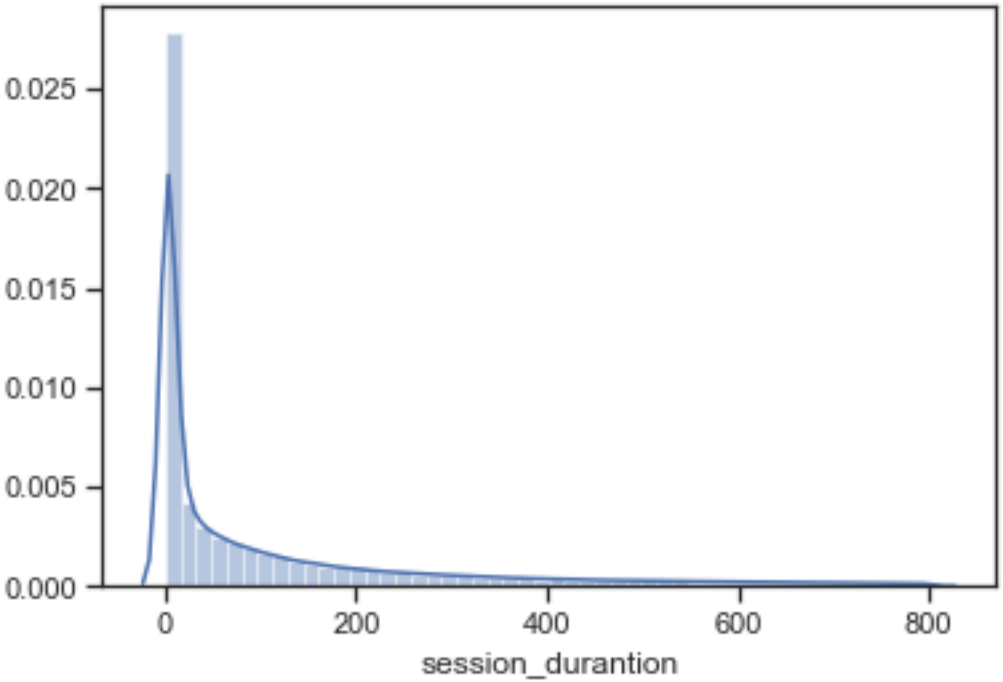
Missing values hit dataset

Data Processing

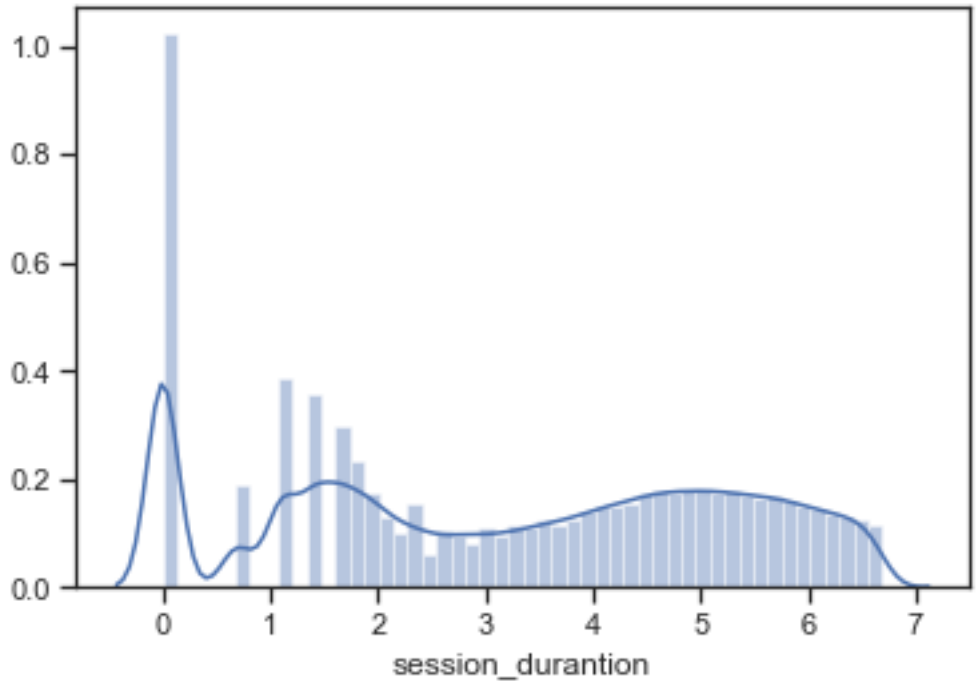
- Handling missing values:
  - session duration: %missing values=7%
  - box plot to quickly check outliers:



- Using IQR score to remove outliers
- Check distrubition of session duration after removing outliers:



=> Session duration is very skewness=> using log transfrom to reduce the skewness



- Replace missing values by median of the variable.

Feature engineering

- Day_of_week and hour_of_day do not present cyclic feature well => create new features using sin, cos function which each extend from 0 to 1 and combine to have the nice cyclical characteristics
- Create new feature, namely the number of location from the feature path_id_set
- Using standard transformaton for all features

Training phase

- Training set: 70% of available values hit dataset, test set: missing values hit dataset (since they have similar characteristics), validation set: 30% of available values hit dataset
- Using logistic regression, random forest, Xgboost and neural network to predict (using 10 folds cross validation for hyperparameter tuning)

Results

Model	RMSE train	RMSE test
Linear regression	18.09384	17.20458
Neural network	17.1961	16.2896
Random forest	15.5242	16.0345
XGboost	16.9613	16.0476

=> Best model: Random forest

Further work (if having more time)

- Using more outliers detection methods (zscore, Grubbs′stest)
- Using more methods to treat missing values
- Create more new features based on entry page, locale, etc
- Try other ensemble models (bagging, voting, stacking)

Thank you for your time!

Contact: phuongphamminh171@gmail.com