Google Merchandize Store Revenue Prediction

prepared by Taisia Komissarova

About the data

About the data

Dataset contains the Google Merchandise Store data, which includes transactions, demographic characteristics of visitors, their behavior on the website.

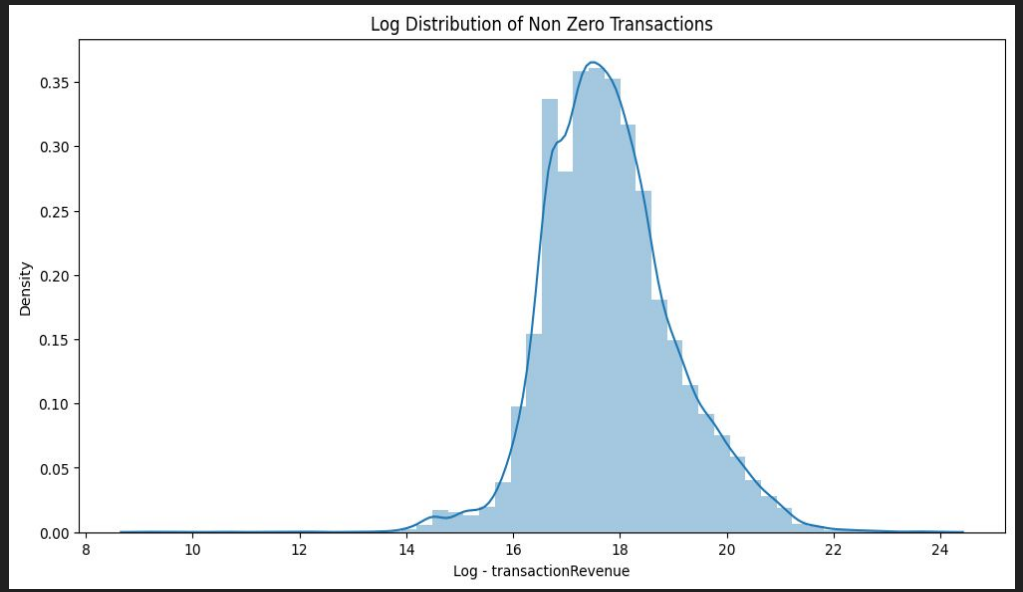
Time interval: 2016-08-01 – 2017-08-01,
which includes train set (70%) from 2016-08-01 to 2017-03-31 and test set (30%) from 2017-04-01 to 2017-08-01.

Number of observations: 903653,
of which 632558 – train set, 271095 – test set.

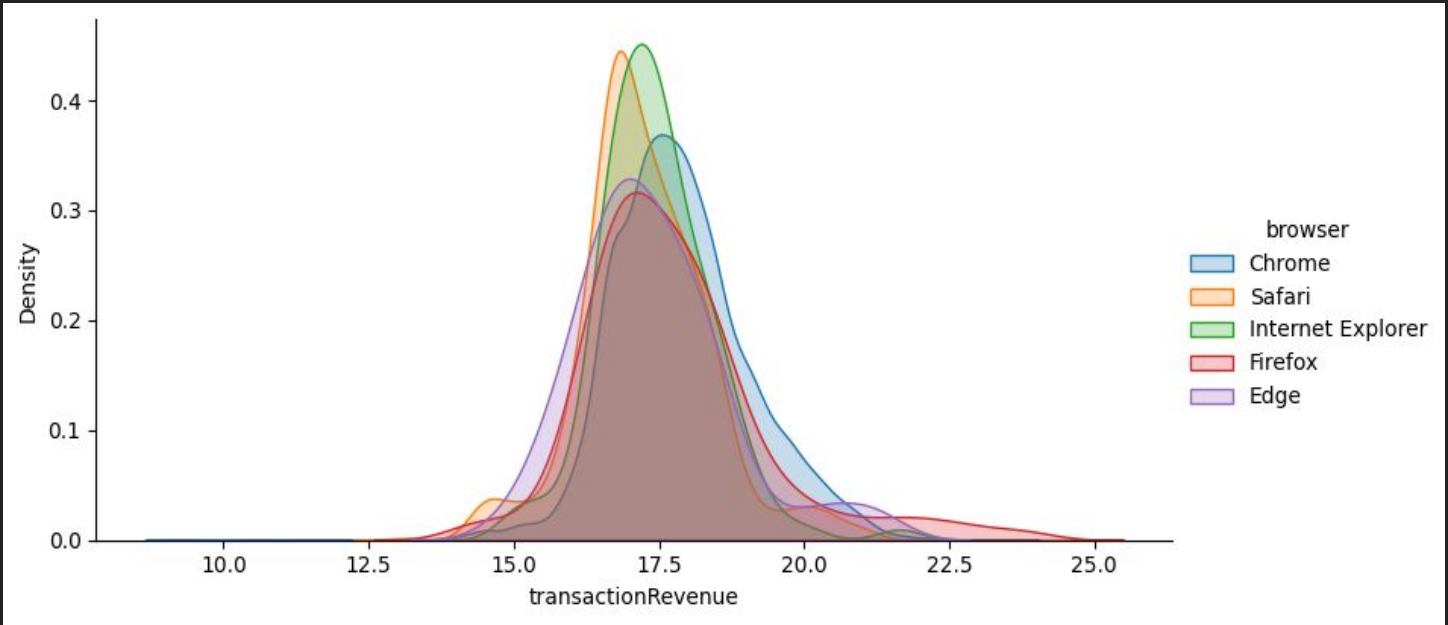
About the data

Target feature transactionRevenue is unbalanced between classes: sessions in which transaction was performed include 11515 observations (1.3% of total).

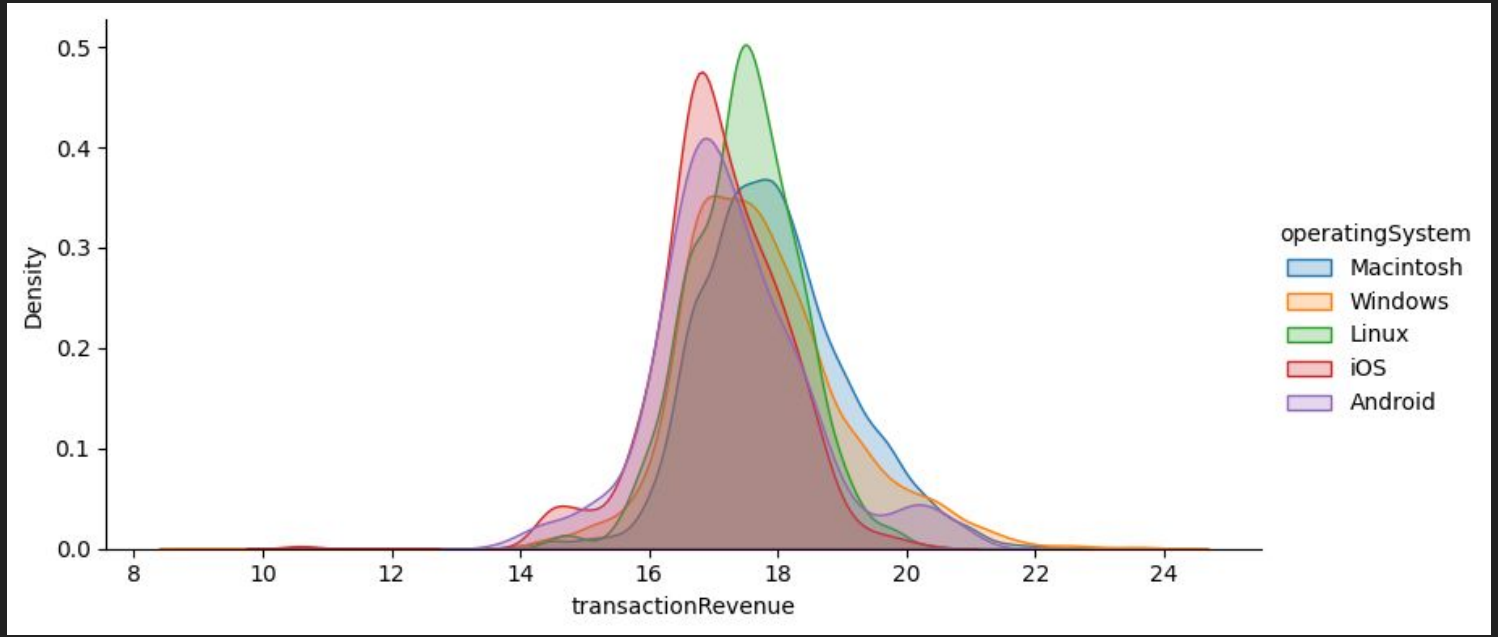
*for convenience, natural logarithm of target feature was taken



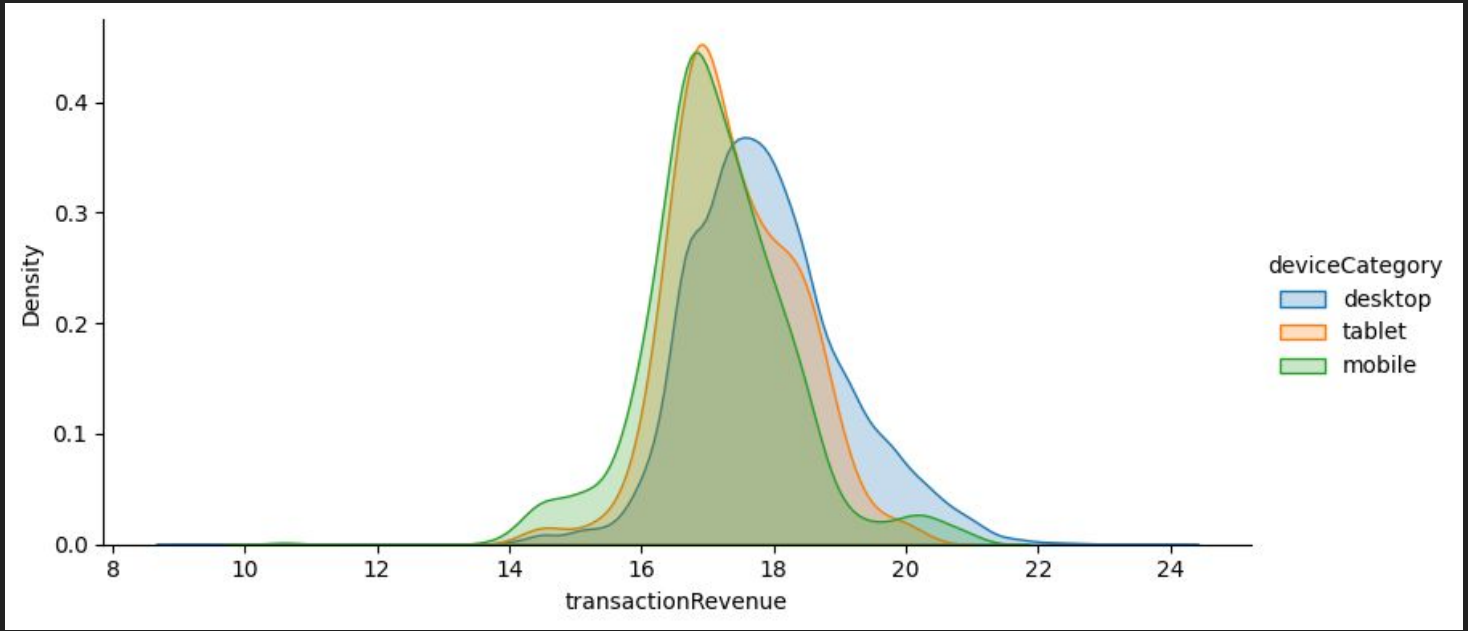
Software insights



Software insights



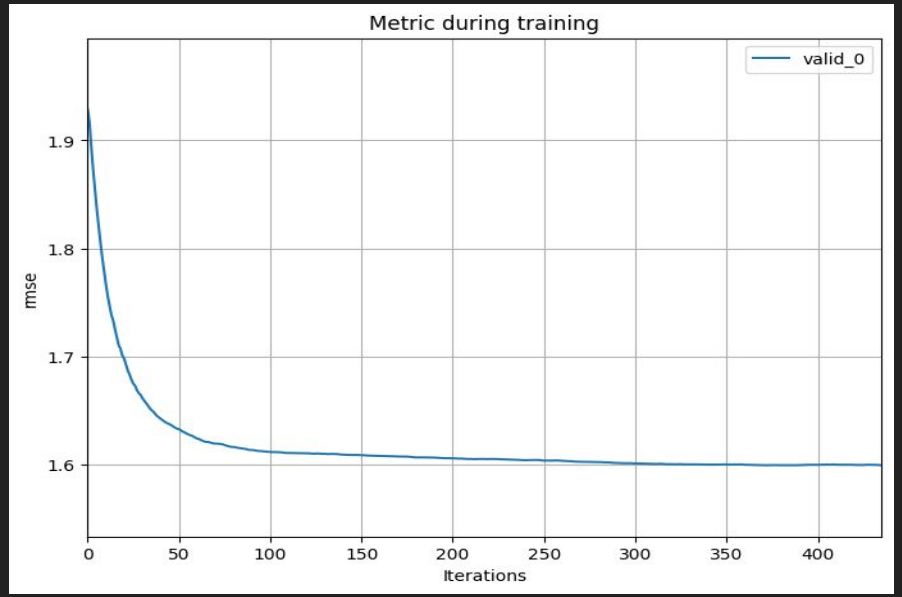
Hardware insights



Model

Model

- LGBMRegressor used to build a model;
- RMSE chosen as a metric;
- learning rate 0.05;
- the best iteration is rmse: 1.59948.



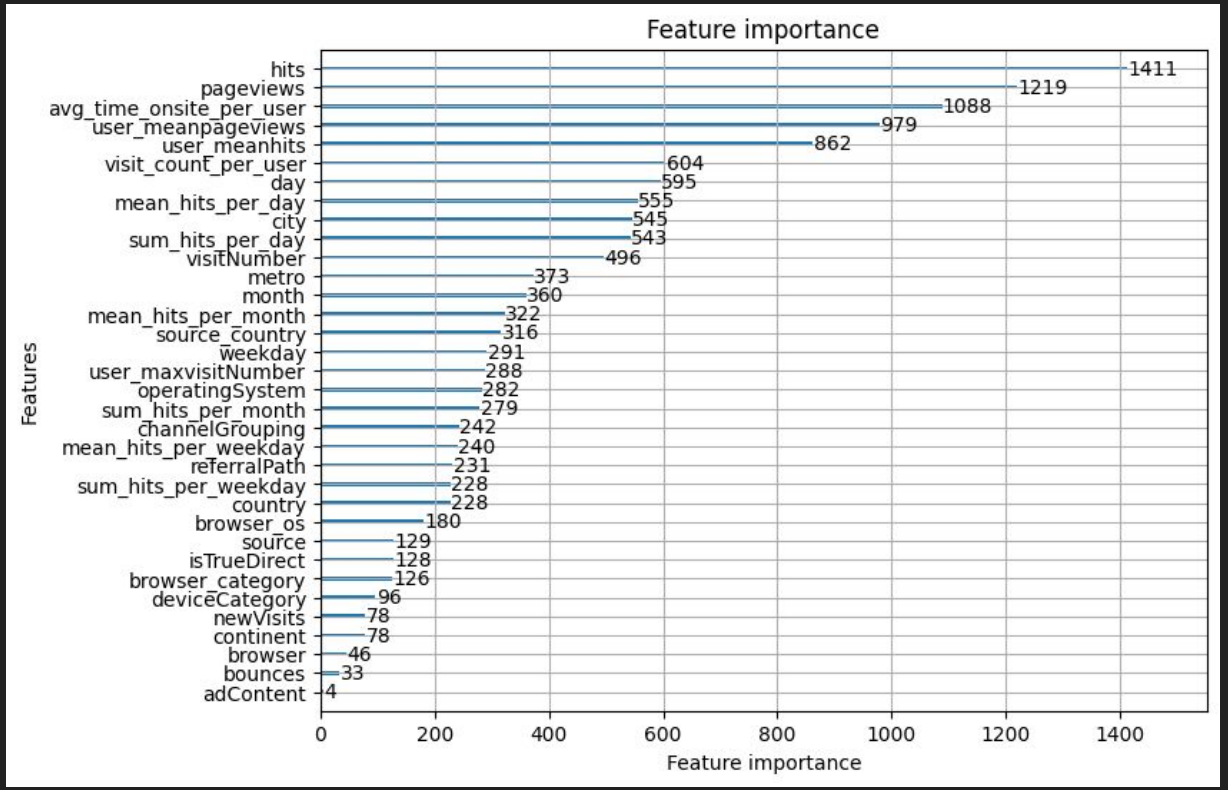
Feature Importance

Feature Importance

After fitting the model the strongest features were identified.

Top 5 are:

1. hits
2. pageviews
3. avg_time_onsite_per_user
4. user_mean_pageviews
5. user_mean_hits



Feature Importance

hits – total number of hits within the session.

Direct proportionality with the target states that *the more hits during the session are performed, the more likely the visitor will complete transaction.*

Feature Importance

pageviews – total number of pageviews within the session.

High correlation with the target states that t*he more pages were viewed by visitor,* the more he/she is interested in the product => *the higher the probability of purchase within the session.*

Feature Importance

avg_time_onsite_per_user – generated feature, which reflects average time of session on user level(expressed in seconds).

Direct proportionality with the target states that *the greater the average time spent by user on a website, the greater the chance he/she will complete transaction.*

Feature Importance

user_mean_pageviews – generated feature, which reflects average number of pageviews on user level, originates from *pageviews* feature.

Direct proportionality with the target states that *the greater the average number of pages, viewed by the user, the greater the chance he/she will complete transaction*.

Feature Importance

user_mean_hits – generated feature, which reflects average number of hits on user level, originates from *hits*.

Direct proportionality with the target states that *the greater the average number of hits, performed by the user, the greater the chance he/she will will bring in revenue*.

Key points

Key points

- Behavioral characteristics are the best assistants for identification of users who will complete transaction and bring revenue.
- Software and hardware used by visitors during session do not influence transaction revenue significantly.
- The more web-pages user previously viewed,
  - the bigger the quantity of previous sessions for user,
  - the more time user previously spent on the website, the bigger the probability he/she will complete transaction.
- Loyal users who have been using the platform for a period of time are likely to become repeated visitors.

Thank you for attention!