Web Traffic Analysis

Ritvik - Dhupkar

Objectives

Impute missing data for the missing ‘hits rows’

	row_num	locale	day_of_week	hour_of_day	agent_id	entry_page	path_id_set	traffic_type	session_duration	hits
0	1	L1	Saturday	23	1	2100	34308;0;183	1	617	\N
1	2	L3	Sunday	8	9	2113	32131;0	2	0	\N
2	3	L3	Saturday	14	9	2100	34330;0	6	17	\N
3	4	L2	Saturday	14	8	2116	89172;0	3	3	11
4	5	L5	Friday	9	2	2100	31777;0	1	610	67

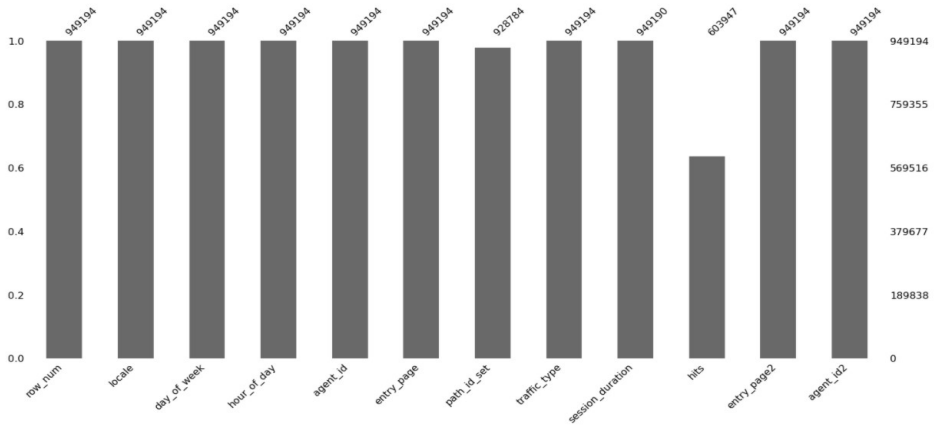
***Web_traffic_analysis_CatBoost_FINAL.ipynb: final CatBoost model (use for evaluation)**

randomforest-and-eda.ipynb : RandomForest model and EDA

Insights on Missing Data

36% of hits data and 2% of session_duration data is missing

Missing Data Percentage Across Columns

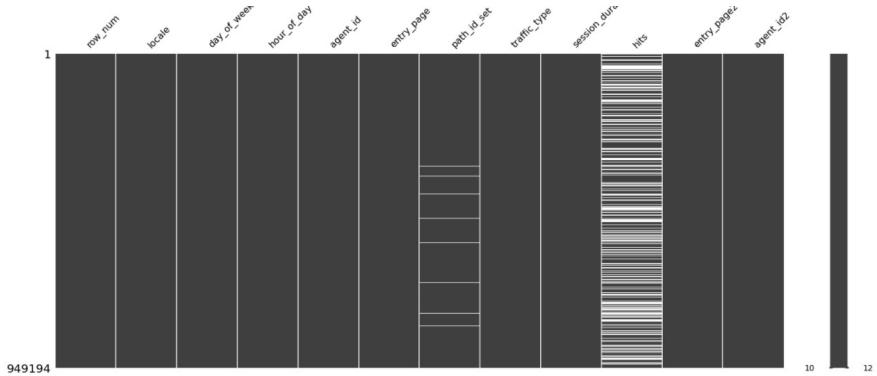


row_num	0.000000
locale	0.000000
day_of_week	0.000000
hour_of_day	0.000000
agent_id	0.000000
entry_page	0.000000
path_id_set	2.150245
traffic_type	0.000000
session_duration	0.000421
hits	36.372649

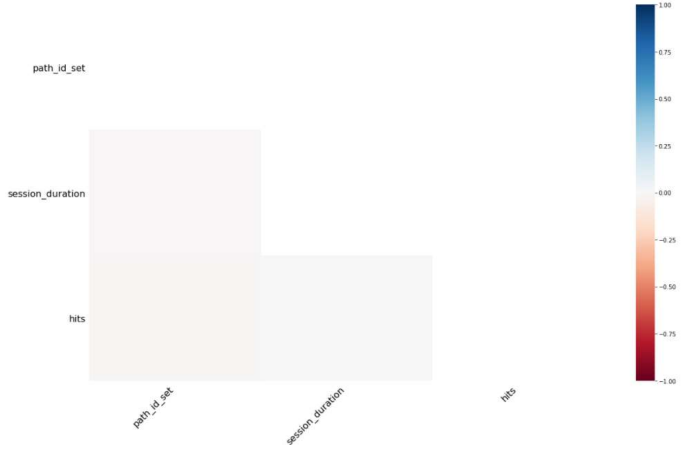
Insights on Missing Data

The missing data is randomly distributed across the domain and there is no correlation in missing Data between columns

Missing Data distribution across columns



No correlation of missing Data across columns



Methodology

Perform Exploratory Data Analysis to understand Factors affecting Hits

Extract features from Web Traffic Information like locale, traffic type, etc.

Clean categorical features: Group low frequency categories of the columns agent_id and entry_page into the new category '999'

Perform One-Hot-Encoding on categorical features

Transform hits and session_duration using log-transform (for Random Forest only) and box-cox transform respectively

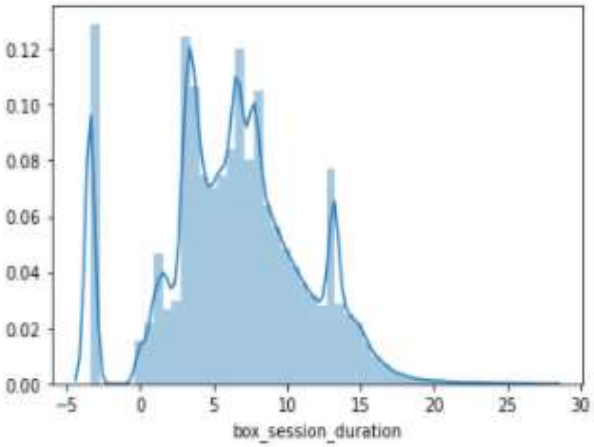
Train a Regression Tree Models to Predict the missing values of 'hits'. The Random Forest Model and Catboost model are used

List of Features

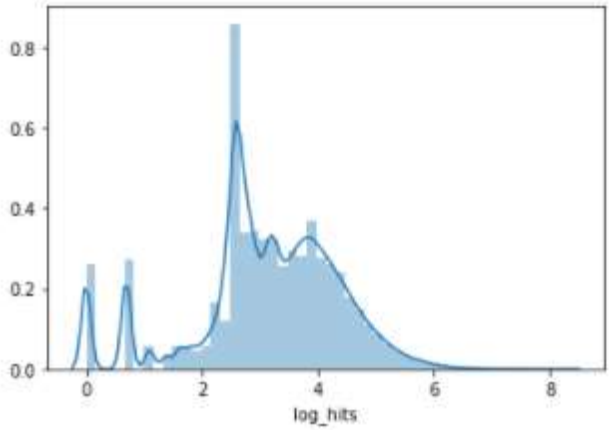
Variable	type	Treatment
row_num	index	index
locale	categorical	one hot encoding
day_of_week	numeric	-
hour_of_day	numeric	-
agent_id	categorical	Group unimportant Categories & one hot encoding
entry_page	categorical	Group unimportant Categories & one hot encoding
path_id_set	categorical	one hot encoding
traffic_type	categorical	one hot encoding
session_duration	numeric	
hits		log_tranform (Random Forest) Catboost (None)

Exploratory Data Analysis on Factors affecting Hits

Box-cox Transformation on session duration



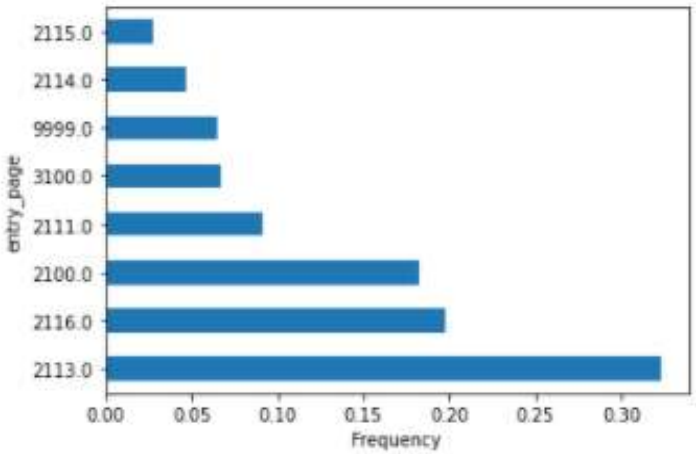
Log Transformation on hits



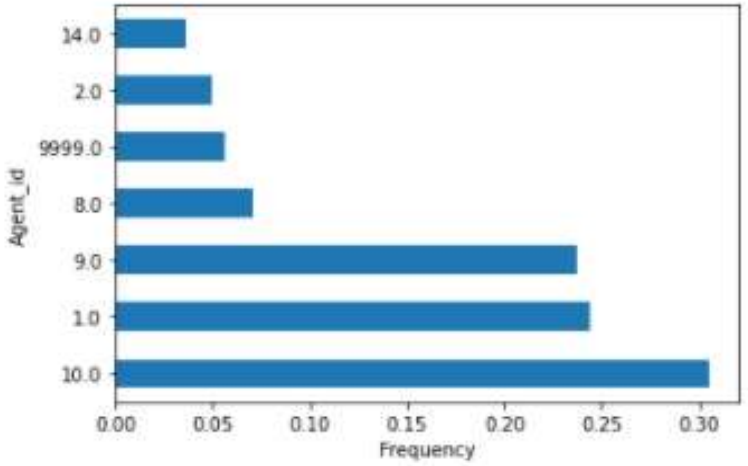
The Session Duration and hits are heavily right tailed, so a box-cox transformation (session_duration) and a log transformation (hits) is applied on them to normalize them,

Transforming categorical variables

entry_page vs frequency



Agent_id vs frequency

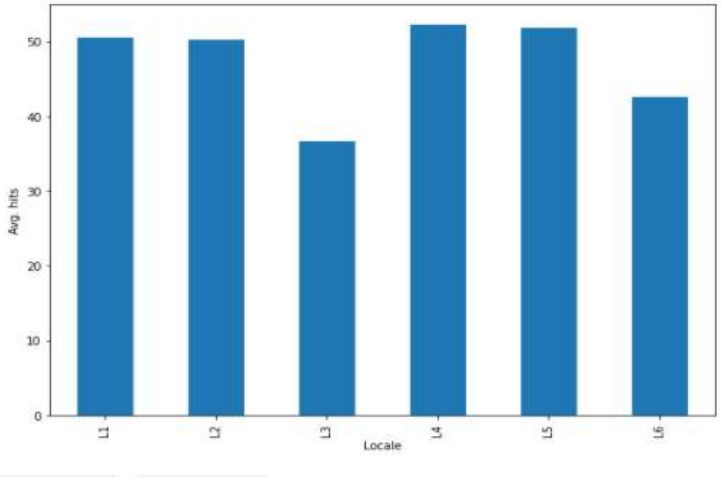


The top 6 most frequently occurring agent_ids and top 7 most frequent entry_page are maintained as categories

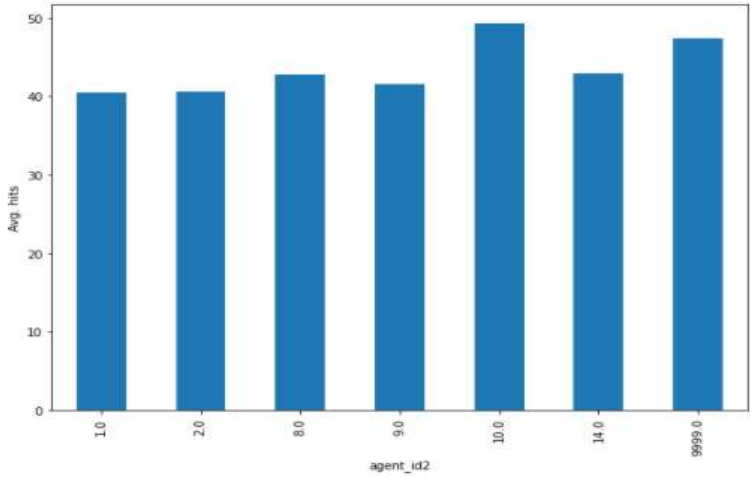
The remaining categories with frequency of 5% are grouped together and labeled ‘999’

Bi-Variate Analysis on hits

Avg. hits vs Locale



Avg. hits vs agent_id



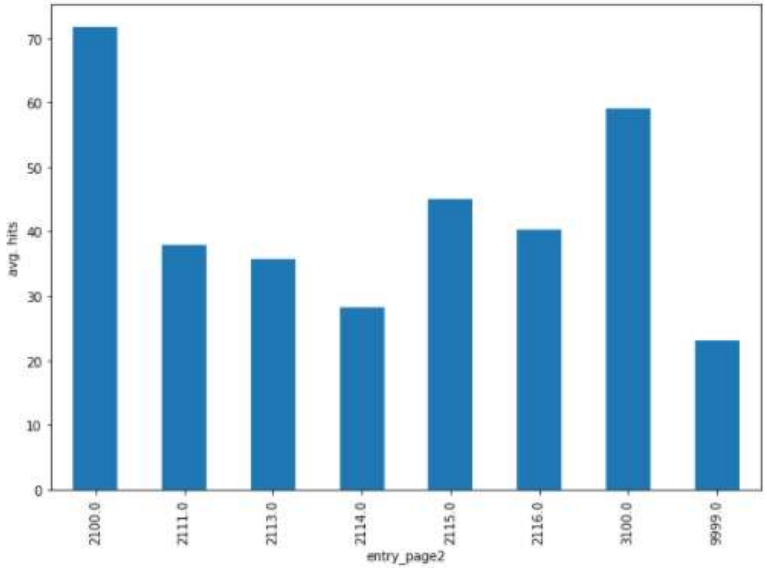
Locale L3 has a significantly lower avg. hits than others

Avg. hits are similar across agent ids

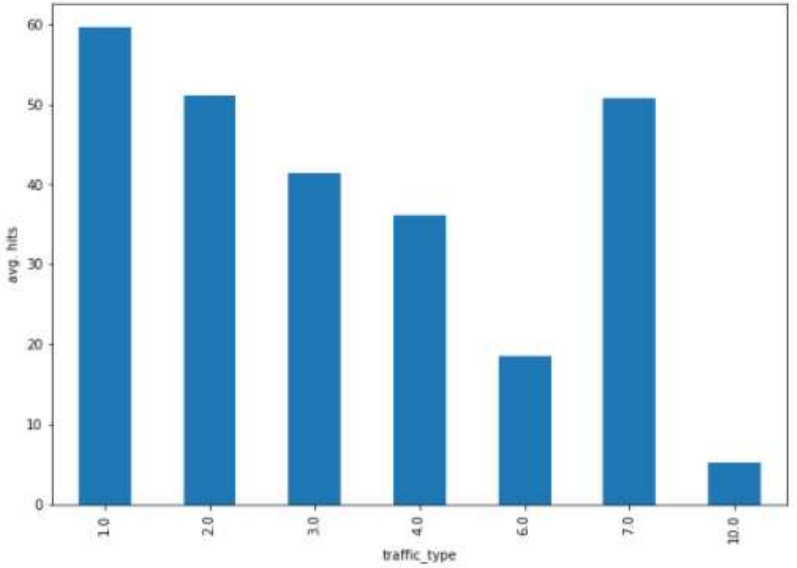
Bottom 5% of least frequently occurring agent ids are grouped together as agent_id = ‘999’

Bi-Variate Analysis on hits

Avg. hits vs entry_page



Avg. hits vs traffic_type

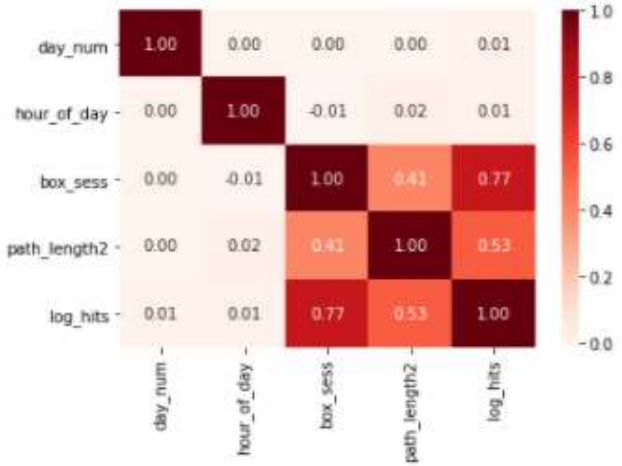


entry_page 2100 has an average hits of 70 while average_hits for other entry pages are lower

Avg. hits differ across traffic types

Bottom 5% of least frequently occurring entry_types are grouped together as entry_type= ‘999’

Correlation of continuous variables with log_hits



Box_sess : box_cox transformation of session_duration

Path_length2 : length of path computed from path_id variable (number of destinations visited)

Path_length and session_duration are highly correlated with log_hits

Random Forest model Results

Methodology

Perform transformations on categorical variables

as listed above. One Hot Encoding is performed on

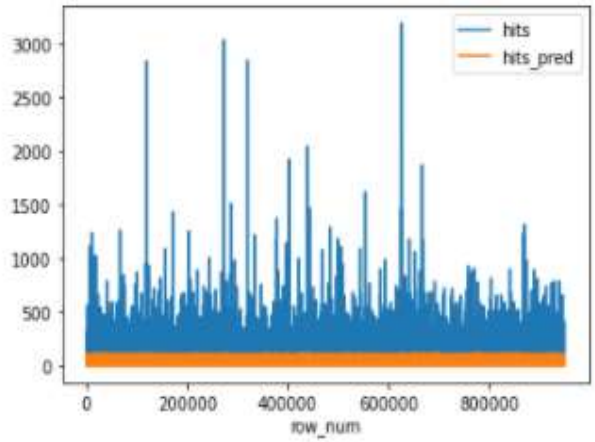
Categorical variables

Transform hits to log_hits. Use log_hits as target variable for training

Take box-cox transform of session_duration

Train the model on available values of hits and perform hyperparameter tuning

Predict the hits on test data with missing values for hits. Take inverse log transform to obtain final predicted values of hits



Validation RMSE = 55.8

hits: Actual values of hits in validation set
Hits_pred: hits predicted by Random Forest model

CatBoost model Results

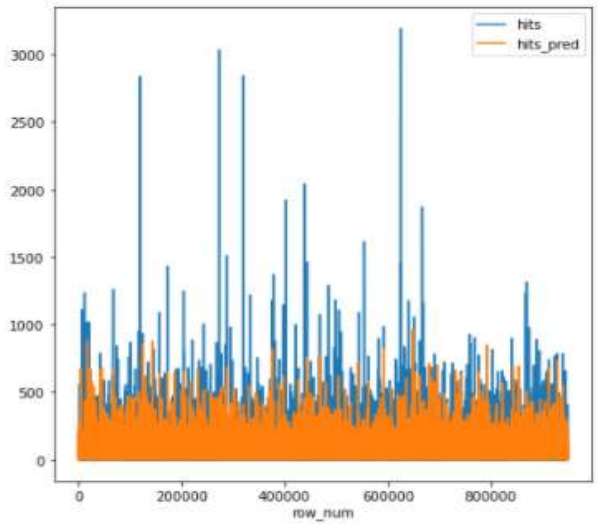
Methodology

From the result plots of the Random Forest, model it is clear that it fails to effectively model ‘count data’

The CatBoost model with objective =‘Poisson’ is used to model the data. This uses a poisson loss function which is used for modelling count data.

One Hot Encoding on categorical variables is performed internally in the CatBoost library

Hits instead of ‘log_hits’ is used as the target variable as a poisson loss function is used



Validation RMSE = 46.4

hits: Actual values of hits in validation set
Hits_pred: hits predicted by CatBoost model

Conclusion

CatBoost model is better able to predict missing platform hits using a poisson loss function

Missing hits: predictions