CUSTORMER REVENUE PREDICTION USING ENSEMBLE METHODS

BY VICTORIA TAYLOR

INTRODUCTION

Kaggle, Google Cloud and Rstudio created the Google Analytics (GA) Customer Prediction competition to demonstrate the business impact that thorough data analysis can have. The data is from the eCommerce Google Merchandise Store (GStore), where Google swag is sold).

The hopeful outcome is more actionable operational changes and a better use of marketing budgets for companies choosing to use the GA platform.

SUMMARY: ABOUT THIS PROJECT

CHALLENGE

Analyze a Google Merchandise Store (also known as GStore, where Google swag is sold) customer dataset to predict revenue per customer.

SOLUTION

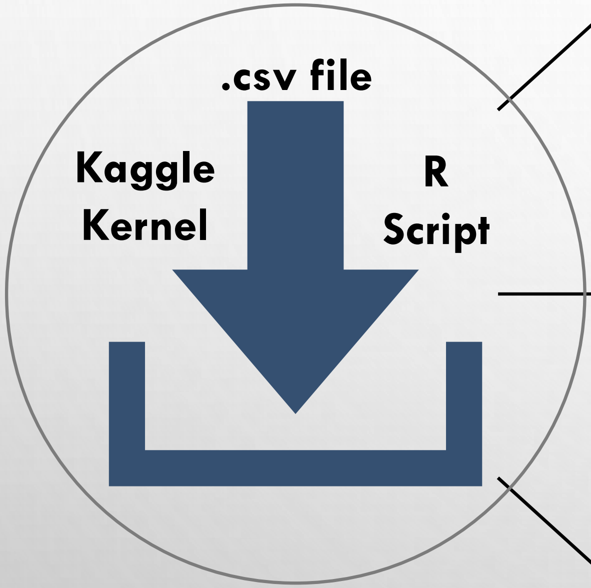
Apply machine learning techniques to web analytics data to build the best possible model for predicting customer revenue.

RESULT

eXtreme Gradient Boosting (XGBoost), one of the best performing regression tree-based ensemble methods was the best model to predict, as best as possible, GStore real-world data.

Using machine learning in addition to web analytics data to predict customer revenue; a way to offer customized service to consumers; maximize revenue; increase loyalty and online presence; appropriate marketing dollars wisely.

THE DATASET



33,818 observations

- 12 attributes including 4 json blobs.
- Variable to predict - "**transactionRevenue**".
- ~ 35% revenue producing observations.
- Evaluation metric – Root Mean Square Error (RMSE)

Limitations

- Deliberate masking of some attributes within json blobs.
- Some attributes were censored to remove target leakage.
- The total impact of the censored masked fields within the dataset is not fully known. Should those fields become available, the models created in this project should be re-evaluated.

Research Questions

- Which model more accurately predicts site customer revenue?
- Which predictors must be tracked to ensure the best possible accurate prediction?
- What key factors may heavily impact the accuracy of the predictions.

APPROACH

All work for this project was conducted in Kaggle kernels using R language. R code for this project is in GitHub - https://github.com/vt101/CustomerRevenuePrediction.

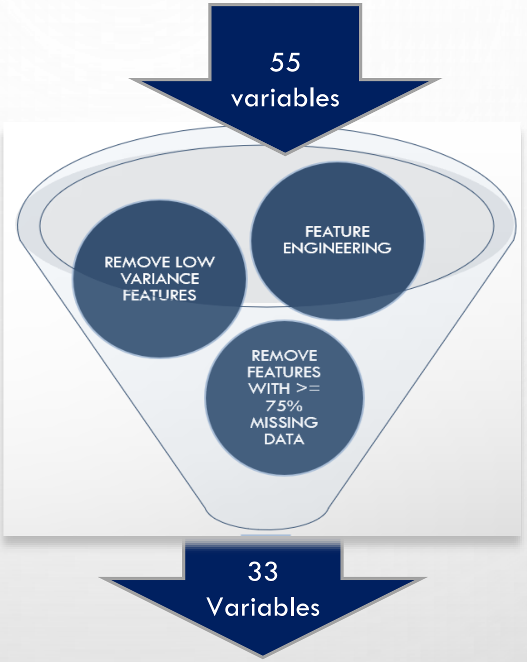
DATA CLEANING & INSPECTION

- import dataset, corrections to data types, etc., redundant features

-initial analysis of data set: dependent variable analysis, distribution, etc.

DATA TRANSFORMATION

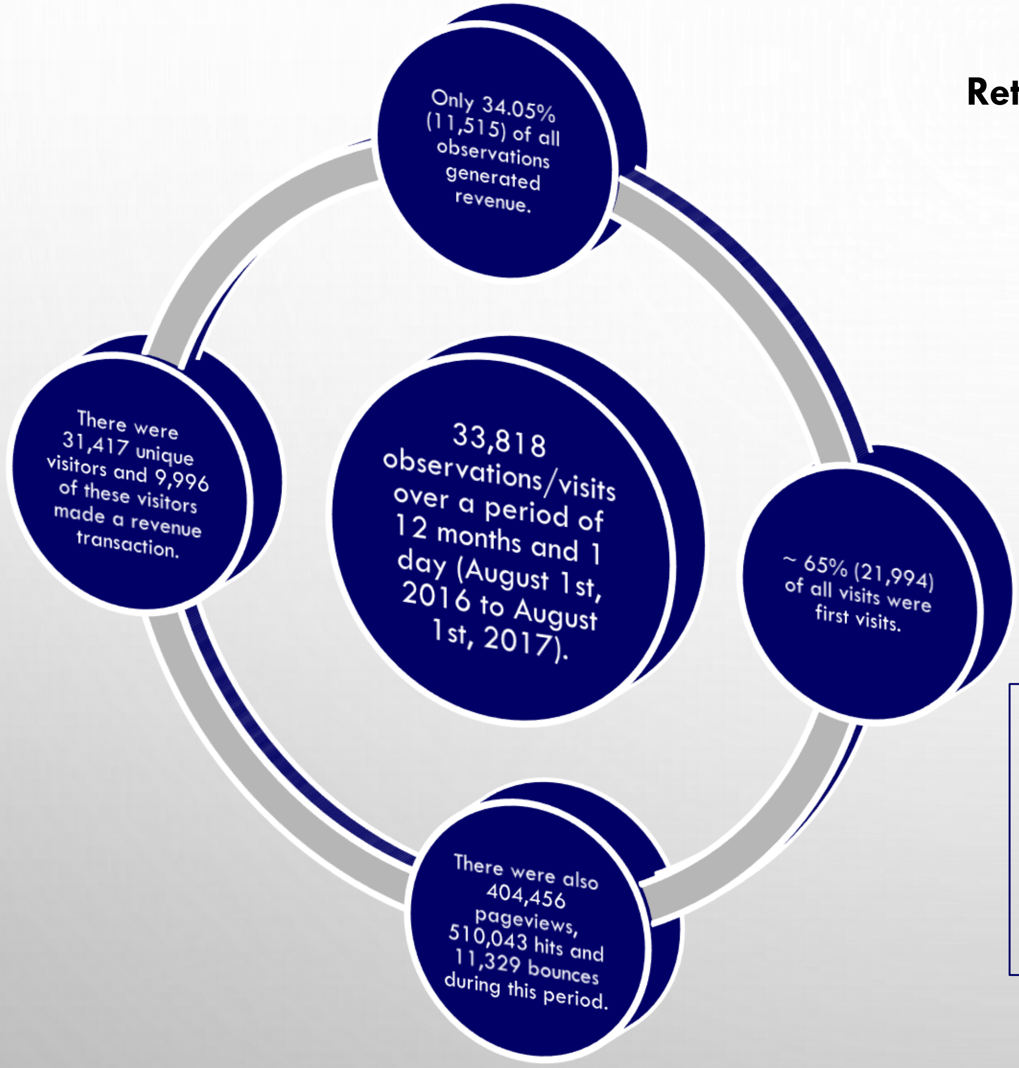
- Feature Engineering
- Correlation matrix
- Prepare data for modeling



BUILD MODELS

1. Random Forest
2. XGBoost

DATA INTERPRETATION

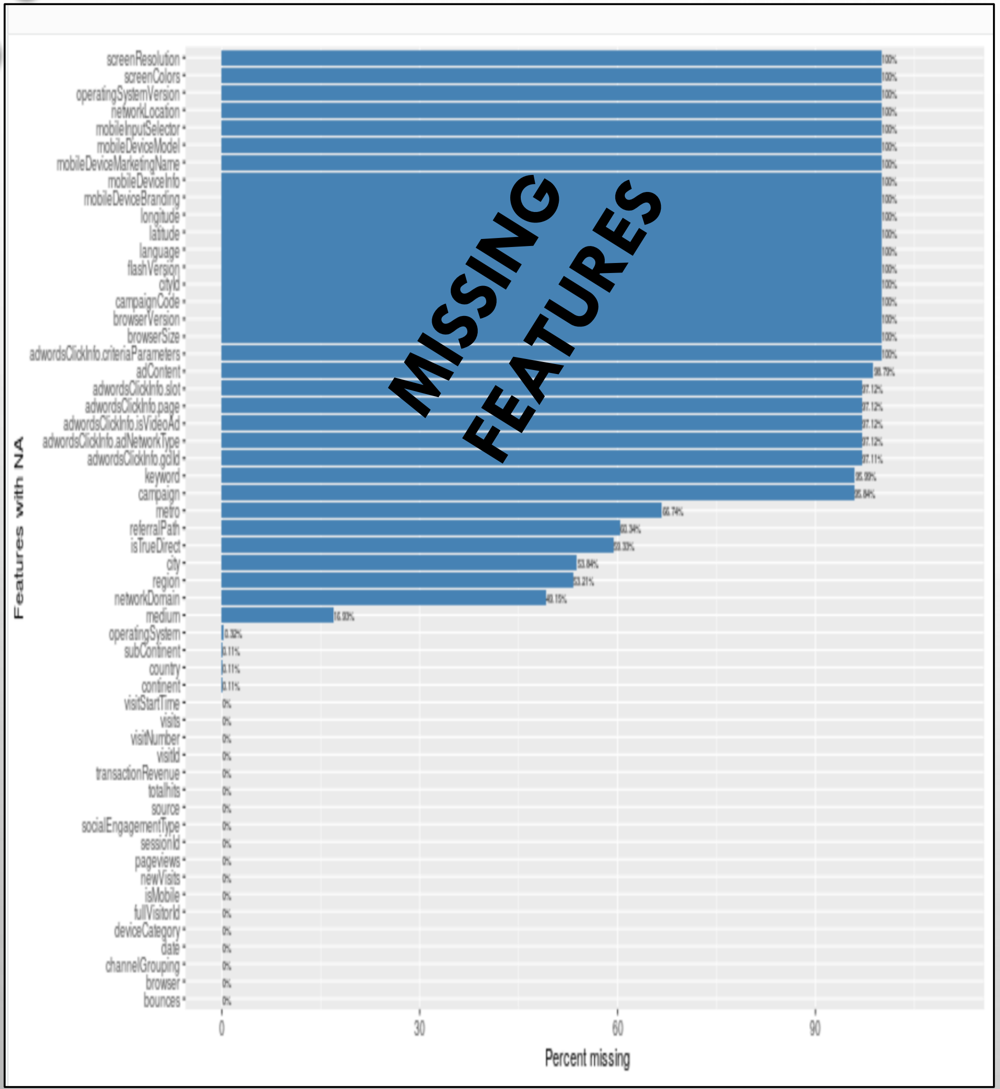


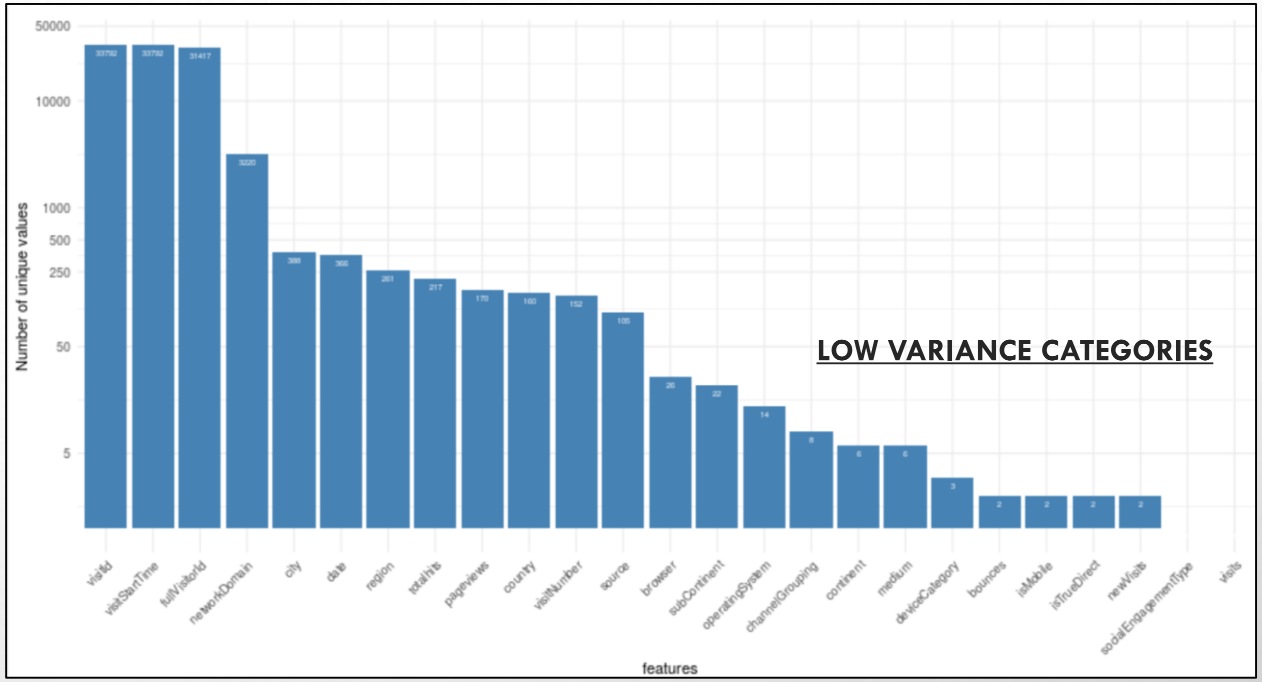
Retention Analysis:

	Visits	Current_visit	Next_visit	retention
1	1 to 2	21994	4716	0.21
2	2 to 3	4716	2268	0.48
3	3 to 4	2268	1333	0.59
4	4 to 5	1333	807	0.61
5	5 to 6	807	550	0.68
6	6 to 7	550	383	0.70
7	7 to 8	383	276	0.72
8	8 to 9	276	231	0.84
9	9 to 10	231	185	0.80
10	10 to 11	185	117	0.63

Of the 65% of first-time visitors to the site only 21% returned for another visit. The volume of these one-time only visits poses a challenge in predicting revenue for each customer to the site. As the number visits increase to the site, the accuracy of the models increases.

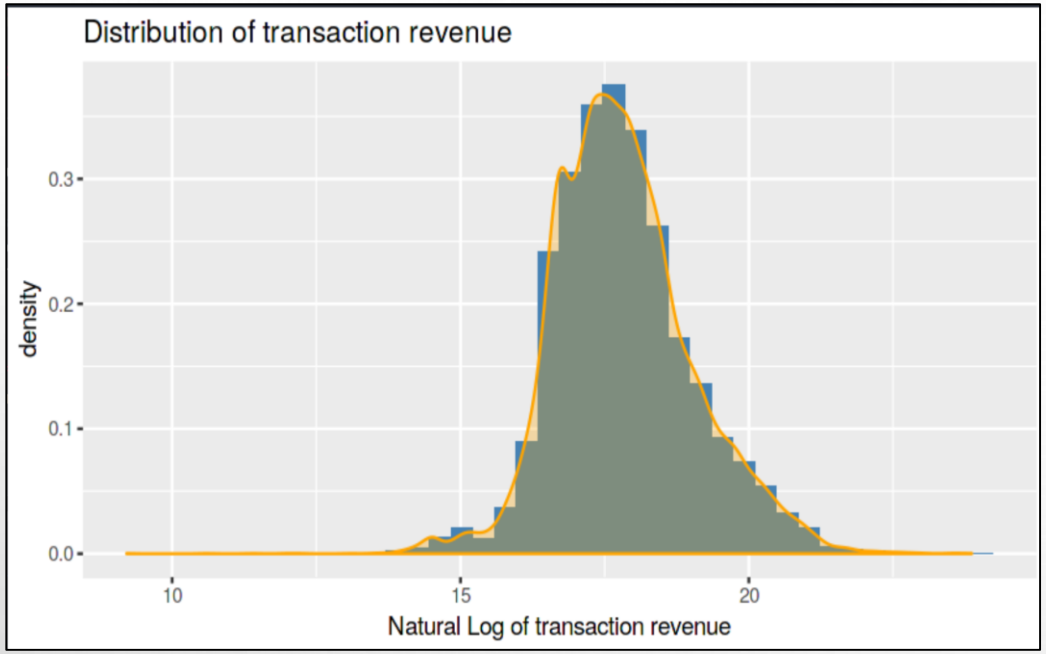
MISSING DATA & LOW VARIANCE





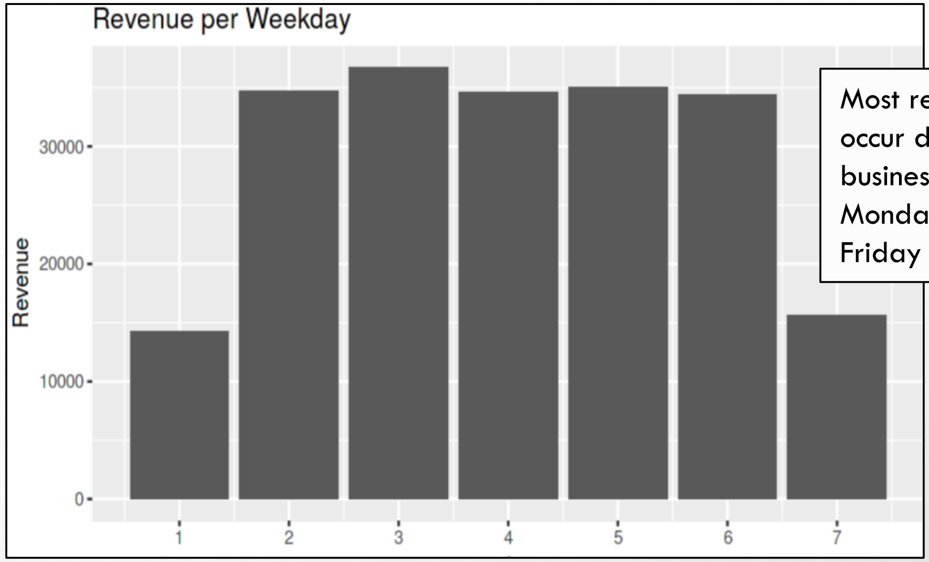
- Data that was missing or blank was coded as “Missing”.
- After transformation, missing data ranged from ~ 0.011% - 0.54%.
  - continent (36), subContinent (36), country (36), operating system (109), region (17996), city (18207), networkDomain (16621) and medium (5724).

DEPENDENT VARIABLE: "transactionRevenue"

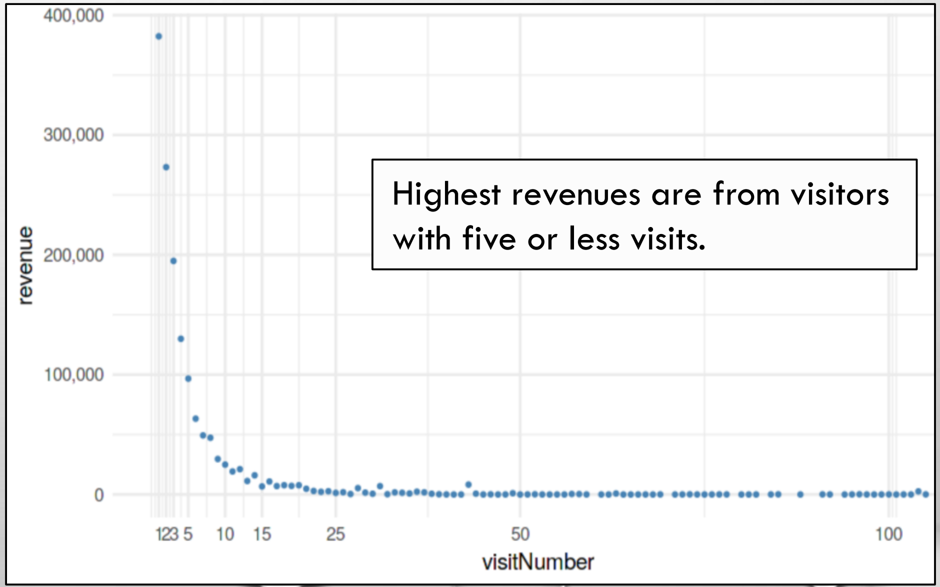


Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
0.00	0.00	0.00	45.54	25.90	23129.50

- The distribution is right skewed.
- ~ 34% of all transactions were revenue producing transactions.
- The values stored were so large the metric was
  - divided by 1e+06 (1000,000) for easy reading and explanation of the metric.
  - Transformed to natural log for prediction purposes.

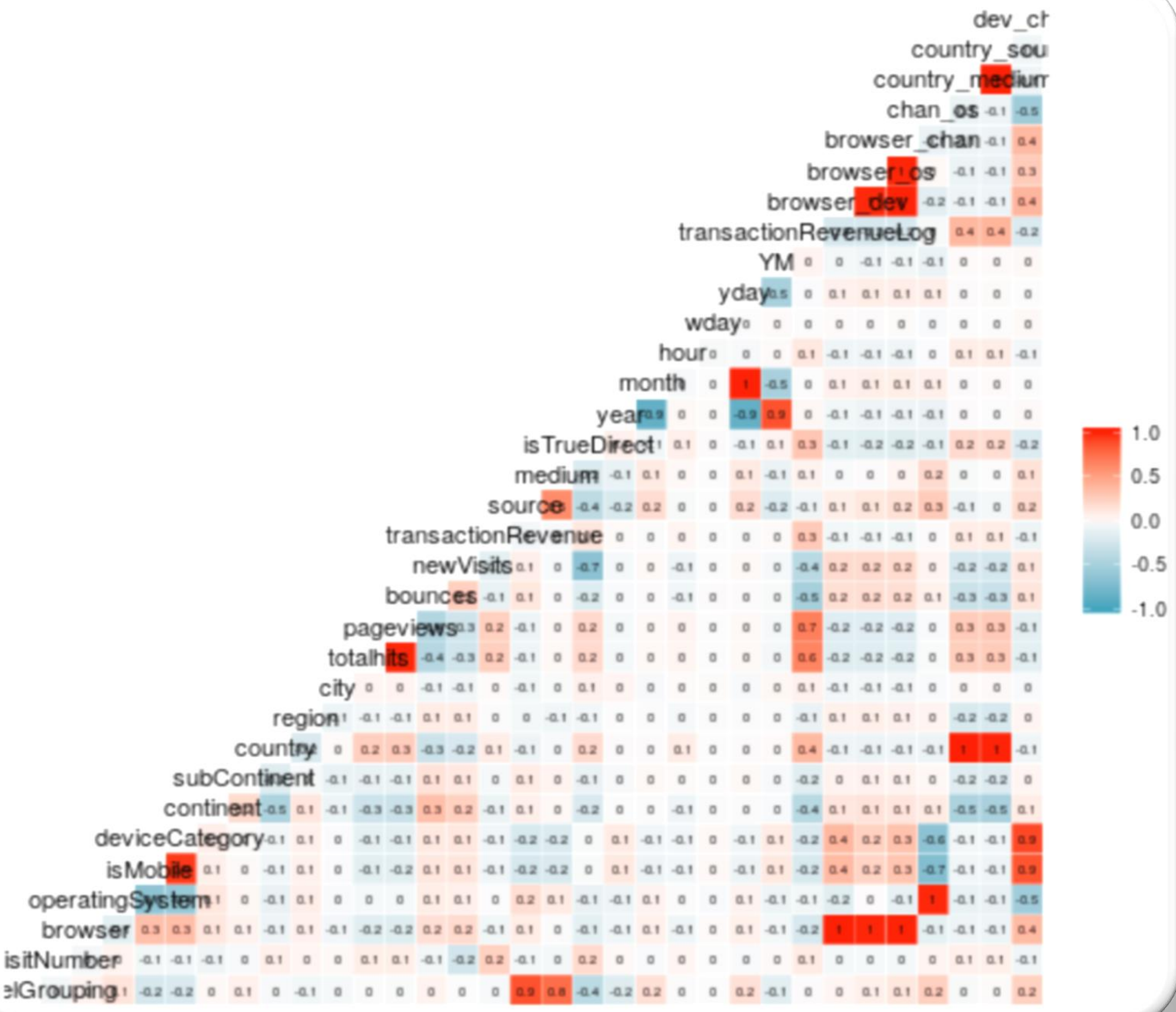


Most revenue occur during business days Monday to Friday (2-6).



Highest revenues are from visitors with five or less visits.

FEATURE CORRELATION



High Correlation

- isMobile to device category(.94);
- channel grouping to source(.87) and medium(.77);
- YM to year(.86), month(.87) and yday(.87);

Medium Correlation

- pageviews to totalhits(-.4)
- operatingSystem to dev_channel (-.5);
- newVisits to medium(-.7);

ENSEMBLE METHODS: MODELS

Random Forest (RF)

Convert Dataset to Numerical Values

- The chosen ensemble methods required data be in a numerical format.
  - i. Converted all non-numerical (categorical) variables to factor.
  - ii. Converted all factor variables to numerical values.

Data Split

- Data was split 65/45, 70/30 and 75/25 for training/testing.
- Using Random Forest model as the benchmark, 70/30 split performed the best.

Cross Validation

- Both methods used the following:
  - i. 5-fold cross validation.
  - ii. Grid search hyperparameter optimization.

eXtreme Gradient Boosting (XGBoost)

The set.seed() function was used when loading the data to ensure results of the evaluation metric were repeatable predictive results.

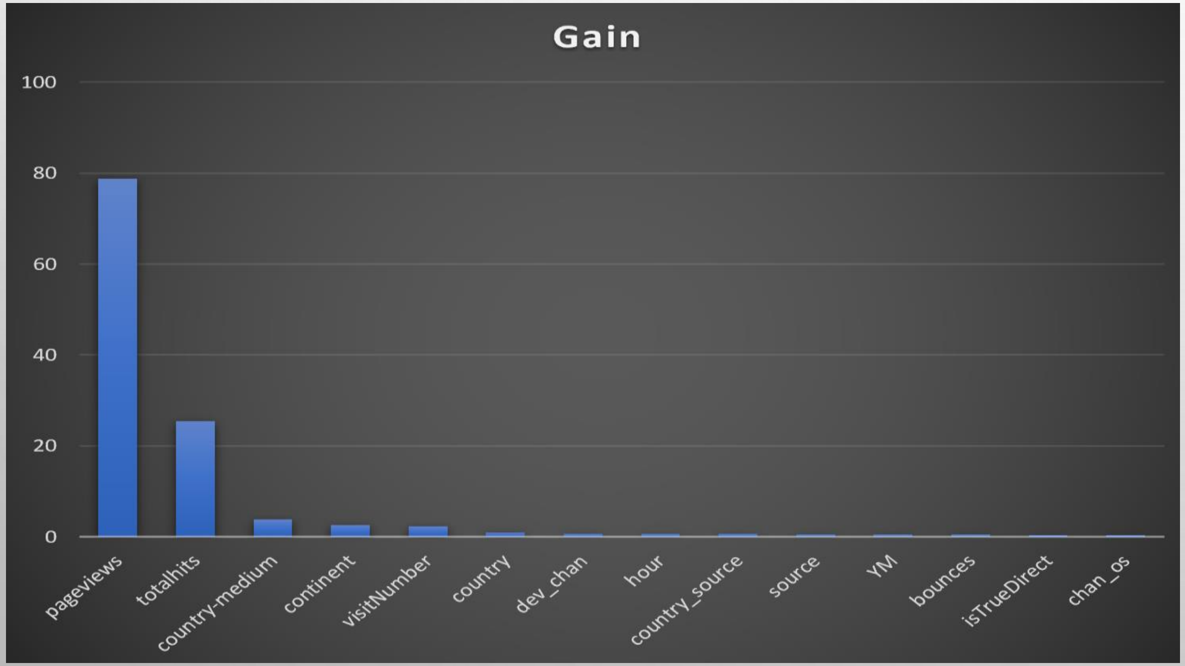
Tree-based ensemble methods

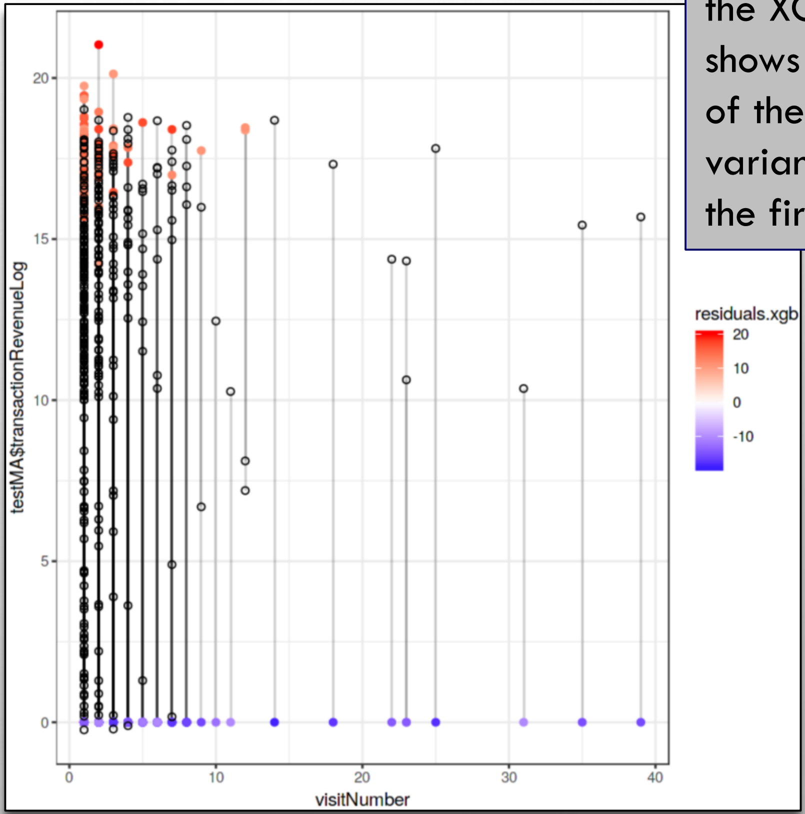
RANDOM FOREST (RF)	• Base model was created with default values. • Hyper-parameter tuning showed error rate stabilized at ~ 150 trees. • Time to execute was 70%+ of XGB execution time. • **Best RMSE: 3.3628**
eXtreme Gradient Boosting (XGBoost)	• Base model was created with default values. • Time to execute was noticeably faster than Random Forest. • Hyper-parameter tuning improved accuracy of prediction. • **Best RMSE: 3.2208**

MODELS: ENSEMBLE METHODS: RESULTS

	Random Forest	XGBoost	% change RMSE
RMSE base model	3.3786	3.2483	-3.87%
RMSE tuned model	3.3628	3.2208	-4.22%
% change RMSE	-0.047%	-0.845%	

Feature Importance of XGBoost Model





A plot of the residual values for the XGB model shows the majority of the largest variances are within the first 5 visits.

CONCLUSIONS

The eXtreme Gradient Boosting (XGB) model is fast, flexible and the better suited predictor of this real world GStore dataset. It will be possible to apply this model to other real-world e-commerce site data.

The top ten most important features to track for most accurate predictions are pageviews, totalhits, country-Medium, continent, visitNumber, country, yday, dev-chan, hour and country-source.

To further improve prediction values:

The impact of censored and/or masked data within the dataset, currently unknown, should be studied.

Additional analysis on customer behaviour during the first 5 visits and how that behaviour is tracked will likely impact prediction accuracy.

Q&A