Google Analytics – Transaction Classification: Which Google Merchandise Store visitors are most likely to make a purchase?

The goal of the project was to predict which Google Merchandise Store customers are the most likely to make a purchase. This would help marketing teams make more appropriate investments in promotional strategies. In the exploratory data analysis, it was clear that while the Store's website users are global, the vast majority of the revenue is concentrated in the US. A few different algorithms that could be suited for classification were considered: logistic regression, linear discriminant analysis and random forest.

Data Exploration

I. Original Variables and Descriptions

1. Full Visitor ID – unique identifier for each user of the Google Merchandise Store. Combination of visitID and sessionID
2. Channel Grouping – an umbrella for traffic source and medium; it is a group of several traffic sources with the same medium. Categories:
    - *Organic Search* – unpaid search results
    - *Social* – visits from social media websites
    - *Direct* – users navigate directly to URL
    - *Referral* – Indicates traffic where users clicked a link from another site, excluding major search engines
3. visitStartTime – the time the user started the session. Helped us obtain information regarding day of the week and month, which we used in exploratory data analysis
4. *Device*
    a. Device Category – type of device from which the visitor accessed the Google Merchandise Store
        - Subcategories: desktop, mobile, tablet
    b. Operating System – OS of the device from which the visitor accessed the Google Merchandise Store
        - Subcategories: Windows, Macintosh, Android, iOS, Linux, Chrome OS, Windows Phone
    c. Browser – browser from which the visitor accessed the Google Merchandise Store
        - Subcategories: Chrome, Safari, Firefox, Internet Explorer, Android Webview, Edge, Samsung Internet, Opera Mini, Safari (in-app), Opera
5. GeoNetwork – geographic specifications. Split into continent, subcontinent, country and city
6. Session ID – unique identifier
7. *Totals*
    a. Page Views – number of times a visitor views the same page on the website
    b. Bounces – a single-page session on the website. The user visits the website and then closes out of the page almost immediately; session duration is zero seconds. The greater the bounce rate, the less likely we are to generate revenue. After all, if the visitor is not spending any time on the website, there is no possibility for a transaction to take place. Possible values: 0, 1
    c. Hits1 – an interaction that results in data being sent to Google Analytics. Can be thought of as a "click."
    d. timeOnSite – total time of the session. Units: seconds
    e. newVisits – number of new users in session. If first visit, value is 1.

   f. SessionQualityDim – an estimate of how close a particular session was to resulting in a purchase. Possible values: 1 to 100. The closer it is to 100, the closer that session was to a transaction
      - If value is 0, then Session Quality was not calculated for the session

8. *Traffic Source*
   a. Source – origin of traffic, such as search engine or domain
   b. Medium – category of traffic source
9. *Transaction*
   a. Transactions – total number of transactions within session. Observed values: 0 - 8
   b. totalTransactionRevenue – total transaction revenue
10. visitID – another identification, unique to user. Can be used in conjunction with fullVisitorID for a completely unique identifier
11. visitNumber – session number for user. If first session, value is 1

Classification Algorithm:

I. Random Forest

First, Random Forest, an ensemble method, was run to understand which of the 24 variables were most likely to be predictor variables. It builds multiple decision trees and merges them together to produce a more accurate and stable prediction. It can be used for both classification and regression problems, which form the majority of current machine learning systems. It adds additional randomness to the model, while growing trees. Instead of searching for the most important feature while splitting a node, it searches for the best feature among a random subset of features. This results in a wide diversity that generally results in a better model.

Variable Importances:

	variable	relative_importance	scaled_importance	percentage
1	totalTransactionRevenue	1973904.875000	1.000000	0.369115
2	transactions	1535284.625000	0.777791	0.287094
3	transactionRevenue	1045919.375000	0.529873	0.195584
4	pageviews	361537.187500	0.183158	0.067606
5	hits1	239422.203125	0.121294	0.044771
6	timeOnSite	108548.578125	0.054992	0.020298
7	sessionQualityDim	39511.531250	0.020017	0.007389
8	country	25118.466797	0.012725	0.004697
9	subContinent	10193.271484	0.005164	0.001906
10	channelGrouping	3816.841064	0.001934	0.000714
11	visitNumber	1866.543457	0.000946	0.000349
12	operatingSystem	1209.893311	0.000613	0.000226
13	deviceCategory	407.504333	0.000206	0.000076
14	visitStartTime	284.924774	0.000144	0.000053
15	date	244.047333	0.000124	0.000046
16	visitId	184.152786	0.000093	0.000034
17	isMobile	161.404129	0.000082	0.000030
18	sessionId	41.571072	0.000021	0.000008
19	browser	12.665272	0.000006	0.000002
20	adwordsClickInfo.page	0.904538	0.000000	0.000000

II. Linear Discriminant Analysis (LDA) – To find a linear combination of predictors that gives maximum separation between the data centers

First, for several variables that had a significant amount of missing values, a null value was assigned. For reference, below is a confusion matrix that shows the number of transactions from the train data (0 = no transaction; 1 = transaction). Percentage of class 1 = 1.06%

0	1
593602	6398

This imbalanced class distribution, which can result in biased and inaccurate models was corrected by down-sampling the majority class and keeping all the rare events.

III. Logistic Regression –The goal is to find the most parsimonious model to describe the relationship between the outcome and predictor variables. This is done by predicting the probability of the dependent variable falling into the category of interest.

Results & Evaluation:

I. Results

1. Random Forest – below is the output

```
> summary(rf)
Model Details:
==============

H2OBinomialModel: drf
Model Key:  DRF_model_R_1543190933615_1346
Model Summary:
  number_of_trees number_of_internal_trees model_size_in_bytes min_depth max_depth
1            1000                    1000             27539598        23        47
  mean_depth min_leaves max_leaves mean_leaves
1   33.60300        956       2449  2187.37300

H2OBinomialMetrics: drf
** Reported on training data. **
** Metrics reported on Out-Of-Bag training samples **

MSE:  0.02988915
RMSE:  0.1728848
LogLoss:  0.09595183
Mean Per-Class Error:  0.06656226
AUC:  0.9832012
Gini:  0.9664023
```

```
Variable Importances:
           variable relative_importance scaled_importance percentage
1         pageviews      1571800.000000          1.000000   0.352731
2              hits1      1242985.125000          0.790804   0.278941
3        timeOnSite       992818.500000          0.631644   0.222800
4 sessionQualityDim       486433.000000          0.309475   0.109161
5       visitNumber       162053.593750          0.103101   0.036367
```

Observations and Interpretations

1. The top predictors of whether a session will result in a transaction are: *pageViews, hits1, timeOnSite, sessionQualityDim, and visitNumber*.
2. Among the top five predictors, *pageViews* and *hits1* contribute more than 60% towards making the determination of whether a transaction will occur
3. The Area Under the Curve (AUC) is 0.983. The closer this value is to 1, the better the model is at discriminating between sessions that result in a transaction versus those that do not.

2. Linear Discriminant Analysis – below is the output

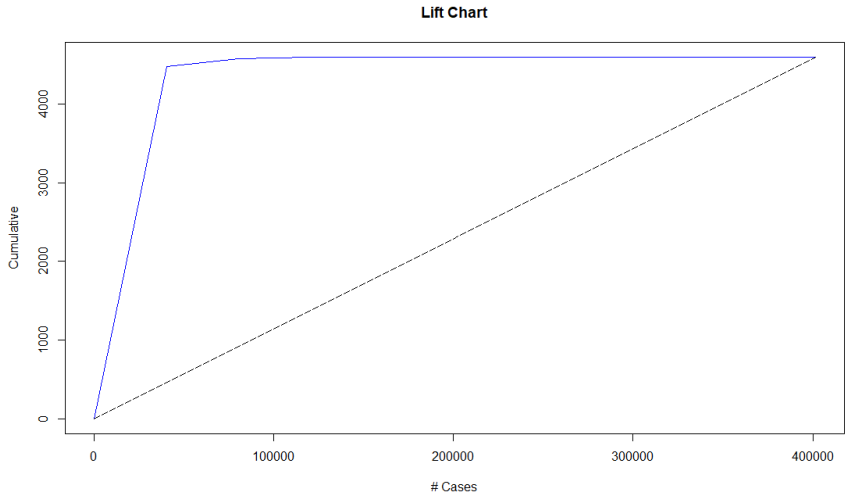
```
Group means:
  pageviews     hits1 timeOnSite sessionQualityDim channelGroupingAffiliates channelGroupingDirect
0  3.404807  4.035561   112.2395          1.481018              0.0198947368             0.1603509
1 26.949359 35.066896  1010.3123         22.653485              0.0003125977             0.1839637
  channelGroupingDisplay channelGroupingOrganic Search channelGroupingPaid Search channelGroupingReferral
0            0.028403509                     0.4227719                0.02629825               0.1184386
1            0.009377931                     0.2966552                0.04032510               0.4599875
  channelGroupingSocial deviceCategorymobile deviceCategorytablet visitNumber newVisits   bounces
0           0.223736842           0.26724561           0.03722807    2.276930 0.7745789 0.5159123
1           0.009377931           0.07361676           0.01500469    3.797124 0.4068459 0.0000000
```

```
Coefficients of linear discriminants:
                                        LD1
pageviews                      0.1826054816
hits1                         -0.0719479395
timeOnSite                     0.0002333377
sessionQualityDim              0.0323943468
channelGroupingAffiliates     -0.2342043419
channelGroupingDirect          0.1769316586
channelGroupingDisplay         0.0974189618
channelGroupingOrganic Search -0.0036501467
channelGroupingPaid Search     0.1822167909
channelGroupingReferral        0.4867919823
channelGroupingSocial         -0.0148075409
deviceCategorymobile          -0.1880623387
deviceCategorytablet          -0.1822372418
visitNumber                   -0.0044714089
newVisits                     -0.3820077256
bounces                       -0.0345583412
```

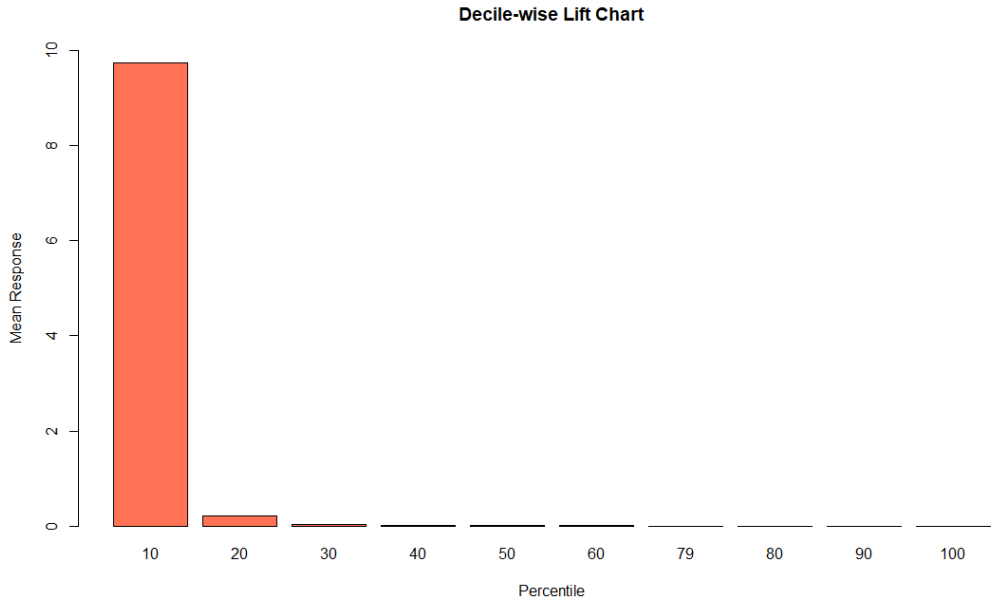
Observations and Interpretations:

1. There is quite a gap between the group means for each of the binary classes for the following variables: pageViews, hits1, timeOnSite, and sessionQualityDim
2. Within the Channel Grouping variable, and amongst all the variables, the referral category has the highest weight for separating the classes.
3. The next variables with the highest coefficients are: newVisits, channelGroupingAffiliates, deviceCategorymobile, and pageViews.
4. This can be interpreted to mean that the above variables are the ones that contribute the most to determining whether a session results in a transaction or not.
5. Specificity: The model correctly classifies a session as not resulting in a transaction 95.8% of the time

Lift Chart – very clearly, the model performs very well in separating the important class from the majority class, as compared to a naïve model



Decile-wise Lift Chart – Taking the top 10% of records that are ranked by the model as most probable to make a transaction (class 1) yields nearly 10 times as many correct 1's as we would have obtained by simply selecting 10% of the records at random.



3. Logistic Regression – below is the output

```
> summary(logit.reg)

Call:
glm(formula = transaction ~ pageviews + hits1 + channelGrouping +
    sessionQualityDim + timeOnSite + visitNumber + deviceCategory,
    family = "binomial", data = Ntrain)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-8.4904  -0.2159  -0.1458  -0.0658   3.2165

Coefficients:
                             Estimate Std. Error z value Pr(>|z|)
(Intercept)                -1.180e+01  8.809e+01  -0.134    0.893
pageviews                   4.168e-01  9.549e-03  43.652  < 2e-16 ***
hits1                      -1.784e-01  6.434e-03 -27.719  < 2e-16 ***
channelGroupingAffiliates   3.277e+00  8.809e+01   0.037    0.970
channelGroupingDirect       7.792e+00  8.809e+01   0.088    0.930
channelGroupingDisplay      7.341e+00  8.809e+01   0.083    0.934
channelGroupingOrganic Search  7.108e+00  8.809e+01   0.081    0.936
channelGroupingPaid Search  7.702e+00  8.809e+01   0.087    0.930
channelGroupingReferral     8.244e+00  8.809e+01   0.094    0.925
channelGroupingSocial       5.406e+00  8.809e+01   0.061    0.951
sessionQualityDim           2.363e-02  9.629e-04  24.537  < 2e-16 ***
timeOnSite                  1.678e-04  4.276e-05   3.926 8.65e-05 ***
visitNumber                -1.346e-03  1.799e-03  -0.748    0.454
deviceCategorymobile       -1.010e+00  6.995e-02 -14.440  < 2e-16 ***
deviceCategorytablet       -1.290e+00  1.652e-01  -7.813 5.57e-15 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 41474  on 63397  degrees of freedom
Residual deviance: 16716  on 63383  degrees of freedom
AIC: 16746

Number of Fisher Scoring iterations: 10
```

Observations and Interpretations:

1. Coefficient of Pageviews: 0.4168. This indicates that a single unit increase in the pageviews, keeping all other predictors constant, is associated with an increase in the odds that the customer will make a transaction by a factor of $e^{0.4168} = 1.517$.
2. Deviance is a measure of goodness of fit of a generalized linear model. Or rather, it’s a measure of badness of fit–higher numbers indicate worse fit. The null deviance shows how well the response variable is predicted by a model that includes only the intercept (grand mean). The model has a value of 41474 on 63397 degrees of freedom. Including the independent variables decreased the deviance to 16716 points on 63383 degrees of freedom, a significant reduction in deviance. In this case, the Residual Deviance has reduced by 24758 with a loss of 14 degrees of freedom.

II. Conclusion and Recommendations

1. Conclusions –

Summary

	Random Forest	Linear Discriminant Analysis	Logistic Regression
Accuracy	95.4%	95.7%	90.3%
Misclassification rate	4.6%	4.3%	9.7%
Sensitivity	87.5%	86.2%	96.5%
Specificity	95.6%	95.8%	90.2%

Using sensitivity as the gold metric, logistic regression is the most robust. The variables that are most important across all three models were *pageViews, hits1, sessionQualityDim*, *timeOnSite*, and *visitNumber.*

2. Recommendations –

My recommendation to the Google Store marketing team is to target their marketing efforts towards users who visit the same page of the store multiple times, as measured by *pageViews* variable. The team should also target those who are actively engaged with the page and spend more time as compared to other users, as measured by *hits1* and *timeonSite*. According to the models, these are the most important predictors in determining whether a session results in a transaction.

It should be noted, again, that most of transactions occurred in the US. Therefore, these predictors were based on this information. We can only make assumptions about what the important predictors are for other countries. Thus, Google Analytics marketing team should focus their efforts on those countries that are generating a high number of sessions, as this indicates potential revenue generation.

Citations:

Ali, Aida, et al. "Classification with Class Imbalance Problem: A Review." *Int. J. Advance Soft Compu. Appl*, vol. 7, no. 3, Nov. 2015, pp. 176–204., home.ijasca.com/data/documents/13IJASCA-070301_Pg176-204_Classification-with-class-imbalance-problem_A-Review.pdf

"BigQuery Export Schema - Analytics Help." *Analytics Help*, Google Analytics, 2018, support.google.com/analytics/answer/3437719?hl=en.

Bruin, Erik. *Google Analytics EDA LightGBM Screenshots*. Kaggle, Oct. 2018, www.kaggle.com/erikbruin/google-analytics-eda-lightgbm-screenshots.

Chatterjee, Sourav. "Discriminant Analysis." BUAN 6356. University of Texas at Dallas, Texas. 6 Nov 2018.

"Google Analytics Customer Revenue Prediction." Accessed October 23, 2018. https://www.kaggle.com/c/ga-customer-revenue-prediction/data

Ottenbacher, Kenneth J., et al. "A Review of Two Journals Found That Articles Using Multivariable Logistic Regression Frequently Did Not Report Commonly Recommended Assumptions." *Journal of Clinical Epidemiology*, vol. 57, no. 11, Nov. 2004, pp. 1147–1152. *ScienceDirect*, doi:10.1016/j.jclinepi.2003.05.003.

Shmueli, Galit, et al. *Data Mining for Business Analytics: Concepts, Techniques, and Applications in R*. 1st ed., John Wiley & Sons, Inc., 2018

Peng, Chao-Ying Joanne, et al. "An Introduction to Logistic Regression Analysis and Reporting." *The Journal of Educational Research*, vol. 96, no. 1, Sept. 2002, pp. 3–14. *ResearchGate*, doi:10.1080/00220670209598786.

Pohar, Maja, et al. "Comparison of Logistic Regression and Linear Discriminant Analysis: A Simulation Study." Vol. 1, no. 1, 2004, pp. 143–161., www.stat-d.si/mz/mz1.1/pohar.pdf