Infer & Predict Gender

John Soler

27 August 2018

This report includes a detailed description of the reasoning and results behind the apprach to answering the posed question. It is meant to serve as documentation for the Data Science Team. For a high overview of the project explained in a simpler way for less technical business users, please see the Powerpoint Presentation.

Contents

1. Setting Up

The first section includes setting the seed for reproducible results, libraries, any required installs, the working directory and producing the password from the passphrase.

Here is a list of libraries used:

```
## Creating r-tensorflow conda environment for TensorFlow installation...
##
## Installation complete.
```

```
##  [1] "jsonlite" "digest"   "dplyr"    "ggplot2"  "pROC"     "caret"
##  [7] "e1071"    "devtools" "cluster"  "stringr"  "gplots"   "tcltk"
## [13] "PRROC"    "class"    "knitr"    "keras"
```

```
## <Tcl>
```

2. Cleaning the data

The question for Stage 2 mentions that there are two corrupted columns. I suspect that these are:

- *days_since_last_order*: This was probably in hours, not in days
- *average_discount_used*: This needed to be divided by 10,000 in order to be used

There were other columns which raised questions:

- There are two columns in the data which are not listed in the description: *redpen discount used* & *coupon discount applied*.
- I noticed that the sum of the individual item types (e.g. *wapp items*, *wftw items*, etc.) do not add up to *items*. I've tried several combinations, also trying to guess if the unisex items are counted twice within the item types.
- There are instances where there are more returns or cancels than orders.
- I see that the columns: *wacc items* and *macc items* are identical.
- The payments columns are binary, yet the description says that these columns should represent the number of orders made with that payment type. I assumed that these are binary columns indicating whether a payment type was used or not.

- The column *used coupons* included some NA values. It would have been ideal to replace these using, for example, a simple regression model based on other columns. However for this purpose I simply replaced the NAs by zeros and made sure to include this column only to crete another more robust feature.

The complete set of checks is presented in the table below.

Summary of data reconciliation tests

test	Pass
duplicate rows	FALSE
wacc_items is not identical to macc_items	FALSE
items by gender add up to total items	TRUE
orders by device add up to total orders	TRUE
orders by address add up to total orders	TRUE
items by type add up to total items	FALSE
there are always less orders than items	TRUE
days since days since first order >= days since last order	TRUE
payments by type add up to total orders	FALSE
every customer has at least one payment method	FALSE
there is no revenue whenever avg discount used is 100%	TRUE
revenues are +ve whenever the avg discount used is < 100%	FALSE
orders > returns	FALSE
orders > cancels	FALSE

In the table above, each test having a “FALSE” means that there is some issue with the reconciliation. I did not investigate each of these into too much detail, but I tried to work around them in the *Feature Engineering and Normalising* section of the code.

3. Inferring the customer gender - Baseline model

A simple baseline model was created to give a first quick labelling of the customers’ gender. This was done to serve as a reference - to make sure that the more complex models to come still produce results in line with intuition.

The Baseline Model chosen is:

If more than 50% of a customer’s purchased items are male items, than the customer is male

The histogram below shows the distribution of the *percentage male* metric, i.e. the number of male items purchased as a percentage of the total items purchased by a customer. There are two things to note here:

- There is a strong polarity in the data, i.e. the good majority of customers either purchase male items **only** or purchase female/unisex items **only**.
- There is an imbalance in the data set - there are more “female only” shoppers than “male only” shoppers. This is important to keep in mind especially when evaluating the Prediction Models later on.

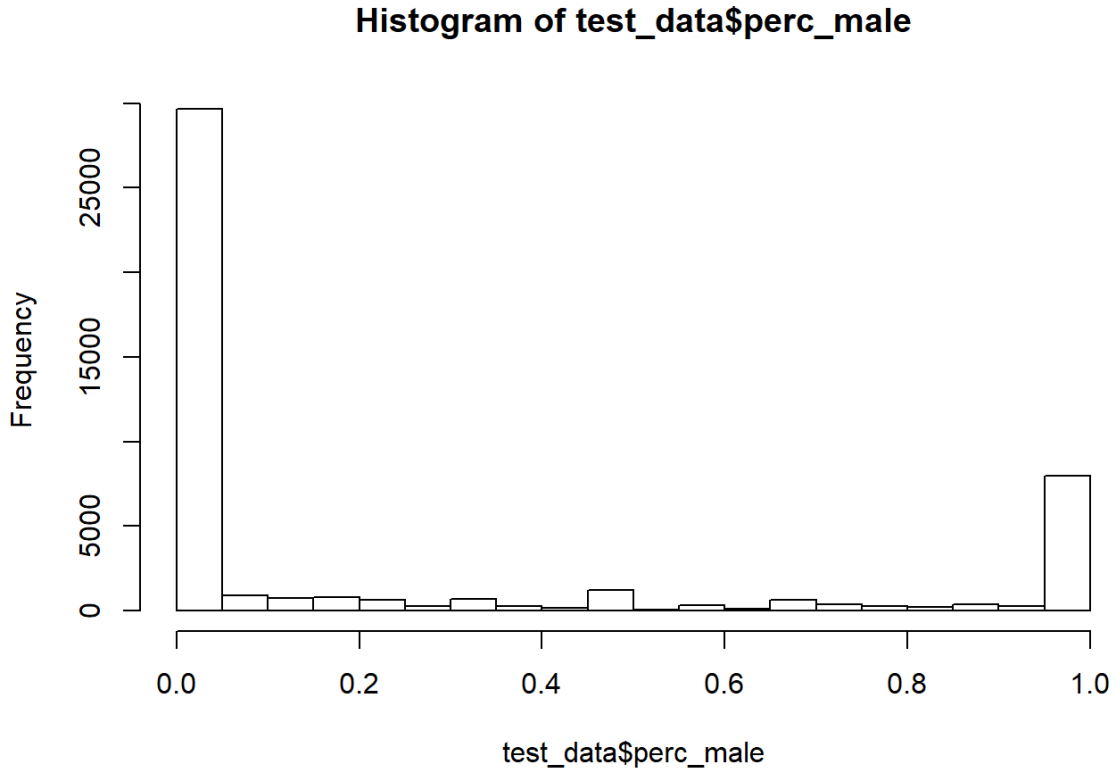


Table of results for the Baseline Model to Infer Gender

isMale_flag_from_Baseline_Model	Number_of_Customers	Percentage_of_Customers
0	35378	0.7685857
1	10652	0.2314143

4. Feature Engineering and Normalisation

Quite a few new features were engineered; mostly ratios such as *revenue per order* or distinct counts such as *distinct count of payment methods used*. Here is a list of all the features including both the provided and the engineered:

```
##  [1] "customer_id"                    "days_since_first_order"
##  [3] "days_since_last_order"          "is_newsletter_subscriber"
##  [5] "orders"                         "items"
##  [7] "cancels"                        "returns"
##  [9] "different_addresses"            "shipping_addresses"
## [11] "devices"                        "vouchers"
## [13] "cc_payments"                    "paypal_payments"
## [15] "afterpay_payments"              "apple_payments"
## [17] "female_items"                   "male_items"
## [19] "unisex_items"                   "wapp_items"
## [21] "wftw_items"                     "mapp_items"
## [23] "wacc_items"                     "macc_items"
## [25] "mftw_items"                     "wspt_items"
## [27] "mspt_items"                     "curvy_items"
## [29] "sacc_items"                     "msite_orders"
## [31] "desktop_orders"                 "android_orders"
## [33] "ios_orders"                     "other_device_orders"
## [35] "work_orders"                    "home_orders"
## [37] "parcelpoint_orders"             "other_collection_orders"
## [39] "redpen_discount_used"           "coupon_discount_applied"
## [41] "average_discount_onoffer"       "average_discount_used"
## [43] "revenue"                        "perc_male"
## [45] "isMale_baseline"                "perc_unisex"
## [47] "days_between_first_last_order" "orders_per_day"
## [49] "items_per_order"                "revenue_per_item"
## [51] "revenue_per_order"              "total_item_types"
## [53] "perc_app"                       "perc_ftw"
## [55] "perc_acc"                       "perc_spt"
## [57] "distinct_item_types"            "perc_cancels"
## [59] "perc_returns"                   "perc_vouchers"
## [61] "multiple_payments"              "perc_mobile"
## [63] "perc_work"                      "perc_home"
## [65] "perc_ppt"                       "perc_oColl"
## [67] "used_coupons"                   "used_redpen"
```

Of these, an initial selection of features was made. At this point, the excluded features were ones which are obviously correlated to another feature, or ones which were considered as not robust enough.

Next, it was made sure that all features were scaled to the range [0,1]. Most of the features were already percentages in the desired range, so those were untouched. For the ones that had larger ranges, a sigmoid function was applied.

Let x be the original column and let y be the normalised column. Then the normalising equation is in the form:

$$y = (1/(1+x))^p$$

where p is selected such that the median of the particular column is normalised to 0.5.

5. Inferring the customer gender - Refined model: Defining Criteria

After having defined the Baseline model and taken key statistics around it, here is an attempt to take that a step further. The idea is to:

1. First select a subset of customers whose gender we can be quite confident of; the **Core Labelled Subset**
2. Use that subset as basis to infer the gender of the rest of the customers

Some quick research was carried out in order to understand better the difference between male and female shopping behaviour. Here are some links which were found useful:

- http://blog.boldmetrics.com/3-differences-between-the-way-men-and-women-shop-for-clothes/ (http://blog.boldmetrics.com/3-differences-between-the-way-men-and-women-shop-for-clothes/)
- https://ecommerce-platforms.com/ecommerce-news/infographic-online-shopping-habits-men-vs-women (https://ecommerce-platforms.com/ecommerce-news/infographic-online-shopping-habits-men-vs-women)

- https://www.paymentsense.co.uk/blog/men-vs-women-online/ (https://www.paymentsense.co.uk/blog/men-vs-women-online/)
- https://www.get.com/blog/infographic-who-rules-online-shopping-men-or-women/ (https://www.get.com/blog/infographic-who-rules-online-shopping-men-or-women/)
- https://medium.com/@rodgerdwightbuyvoets/differences-between-how-men-and-women-do-online-shopping-6e590e54d06f (https://medium.com/@rodgerdwightbuyvoets/differences-between-how-men-and-women-do-online-shopping-6e590e54d06f)

Based on this research it was decided to select the **Core Labelled Subset** using 3 criteria:

1. The percentage of male items from the total items - it is assumed that males tend to purchase more male items
2. A measure of the discounts used - the hypothesis is that females look for discounts more than males do
3. A measure of the order returns/cancellations - the hypothesis is that females return/cancel orders more than males do

Indeed, the table below indicates that there is probably truth in the hypothesis i.e. that females do look for discounts more than males and females do return/cancel orders more than males. Time permitting, statistical tests should be run to support these claims.

Summary of results for the three selected criteria

isMale_baseline	perc_male	test_discount	test_ret_cancel
0	0.0447369	0.5481749	0.1093782
1	0.9366160	0.5094661	0.0642959

The table below shows the exact thresholds for each criteria as well as the number of customers per gender within the Core Labelled Subset. There are roughly 10,000 customers in this subset, which is about 20% of the whole data set.

Core Labelled Subset Summary

Core_Labelled_Subset	perc_male_criteria	discount_measure_criteria	return_cancel_measure_criteria	Number of Customers
Males	= 1	< 0.5	= 0	2318
Females	= 0	>= 0.5	> 0	7701

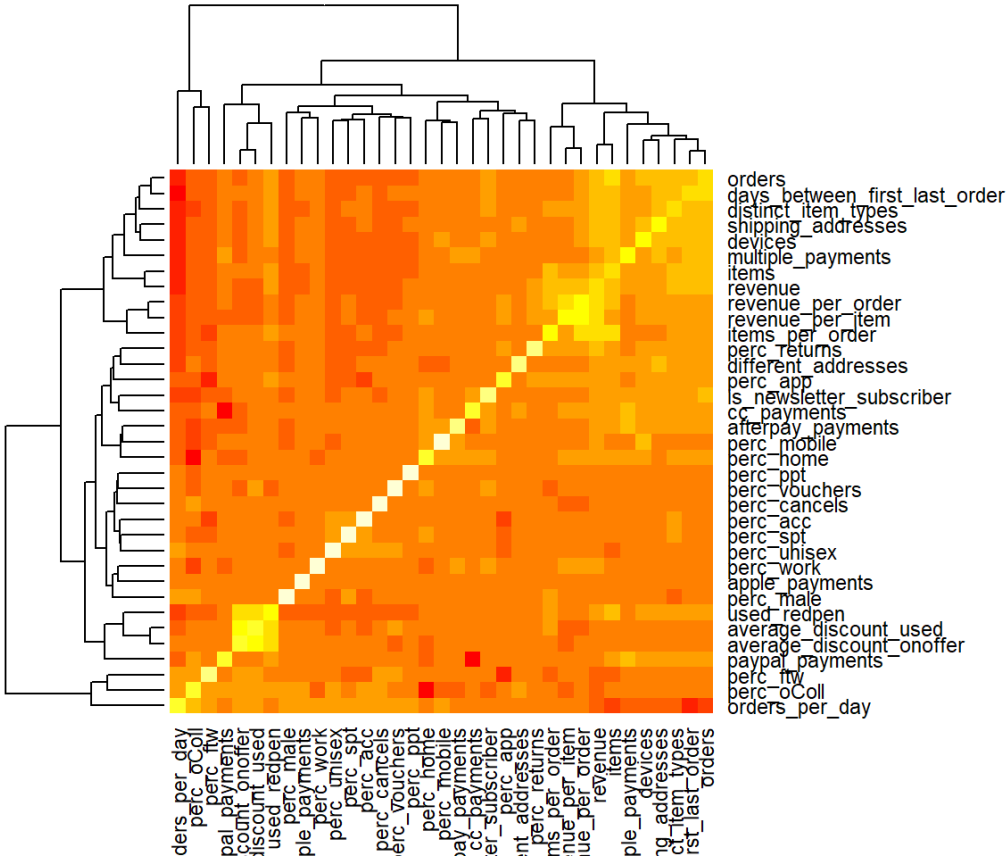
6. Feature Selection

Now we proceed to feature selection. We do this now in anticipation of applying KNN to get the genders of the remaining customers (~80%) in the data set. More details on this in the following section.

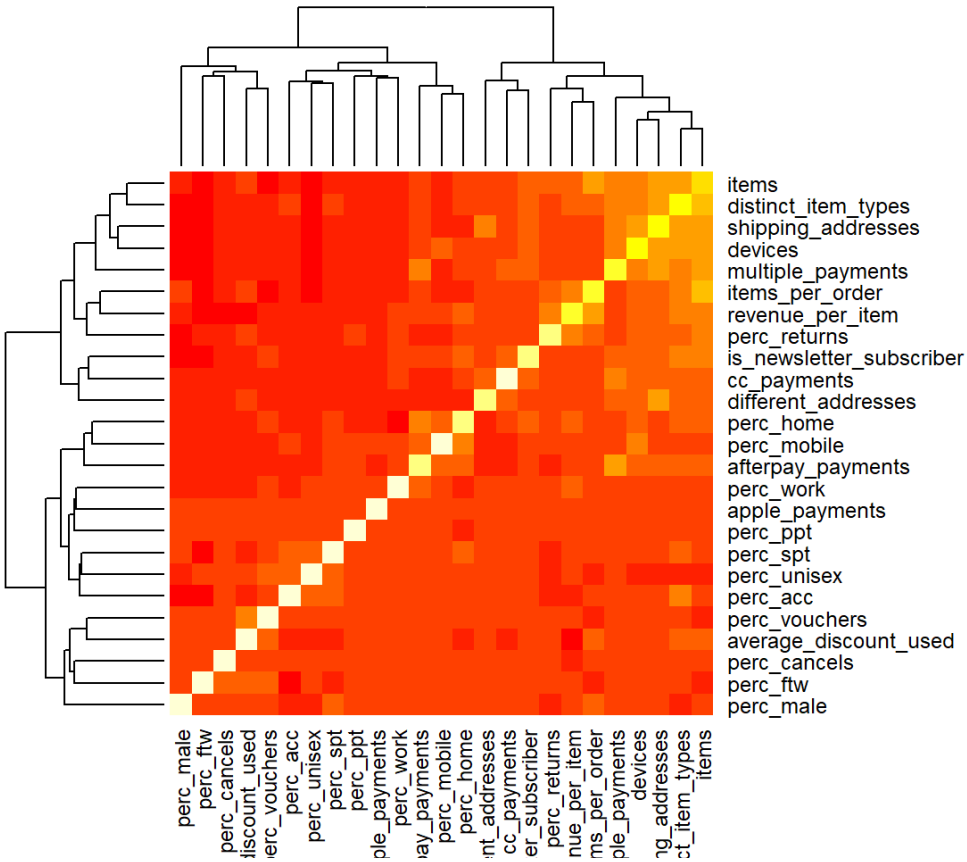
Since time was limited, feature selection (or rather, removal) was done based on the correlation matrix, using both the Pearson and Spearman methods. The correlation matrix was computed and features were removed on the basis of collerations of more than 0.75 (or less than -0.75) and by inspection of the heatmap.

Plotted below are the heatmaps of the correlation matrices before and after feature selection.

Before Feature Selection:



After feature selection:



Here is the final list of selected features to be used for the next stages:

Feature	Description	Meaning_in_real_world
perc_male	Percentage of male items purchased being male items	The degree of preference/need for male items

Feature	Description	Meaning_in_real_world
perc_unisex	Percentage of items purchased being unisex items	The lack of preference/need for gender-specific items
is_newsletter_subscriber	Flag for a newsletter subscriber	Whether or not the customer is willing to hear about new products/promotions
items	Normalised number of items purchased	The Frequency aspect of the customer’s purchase behaviour
different_addresses	Normalised number of times a different billing and shipping address was used	A high number here represents customers who either shop for someone else or are usually not home
shipping_addresses	Normalised number of different shipping addresses used	A high number here represents customers who shop for several other people or their job/lifestyle requires them to move around
devices	Normalised number of unique devices used	A high number here suggests that the customer takes the opportunity to shop during different times of the day and from different locations
cc_payments	Flag for a credit card payment user	Shows customers who trusts the website enough to pay with their credit card
afterpay_payments	Flag for an Afterpay payment user	Could suggest that a customer is tight on budget yet stll willing to purchase exactly what s/he wants
apple_payments	Flag for an apple payment user	Flag for an apple payment user
average_discount_used	Average discount rate of items typically purchased	The degree at which the customer looks for deals - could be either to save money or to buy more expensive items and an affordable price
items_per_order	Normalised number of items purchased per order	A low number here shows that the customer is specific in the way s/he shops - purchasing only an item that is needed at that time
revenue_per_item	Normalised total revenue per item purchased	A high number here shows that the customer prefers higher quality items and has the money for it
perc_ftw	Number of footwear items purchased as percentage of total items purchased	The customer’s interest towards footware
perc_acc	Number of accessory items purchased as percentage of total items purchased	The customer’s interest towards accessories

Feature	Description	Meaning_in_real_world
perc_spt	Number of sport items purchased as percentage of total items purchased	The customer’s interest towards sport itmes
distinct_item_types	Normalised number of distinct item types purchased	The customer’s variety of interest
perc_cancels	Percentage of orders cancelled	A high number here could suggest that the customer is shopping on impulse and easily changes his/her mind
perc_returns	Percentage of orders returned	A high number here could suggest that the customer is either not attentive to detail when shopping, or that s/he likes to try on a range items/sizes before s/he makes a final decision
perc_vouchers	Number of vouchers used as percentage of total items purchased	A high number here suggests that the customer is possibly even willing to wait for a voucher before s/he shops. A low number suggests that the customer is more interested in purchasing what s/he wants at the time that s/he wants it, with less concern about the price
multiple_payments	Normalised number of different payment methods used	Normalised number of different payment methods used
perc_mobile	Percentage of orders via mobile (Msite, Android and iOS)	A high number here may suggest that the customer enjoys shopping on the move. A low number suggests that more shopping is done primarily on desktop, so requireing more attention and possibly with several tabs opened. Alternatively it could simply reflect the type of device available to the user
perc_work	Percentage of orders delivered to work place	A high percentage here suggests that the customer has a job and there is no one home to accept the delivery
perc_home	Percentage of orders delivered to home	A high percentage here suggests that the customer is typically at home or there is someone home to accept the delivery
perc_ppt	Percentage of orders delivered to a parcel point	A high percentage here suggests that the customer is typically not at home and not in a fixed office

7. Inferring the customer gender - Refined model: Applying KNN

The K-Nearest-Neighbours (KNN) algorithm was chosen to propagate the inferred gender from the Core Labelled Subset. This algorithm, being very straight forward and easy to explain, was considered as a good option to go with - it gives practically full control and transpacency about the inference methodology.

From the 10,000 customers in the Core Labelled Subset, 10 samples of 3,000 at a time where chosen on which to apply KNN. It was made sure that each smaple contained a balanced amount of males and females from the Core Labelled Subset. The “K” was also varied from 1 to 5 and the average was calculated at the end, so make sure that the algorythm is not sensitive to any particular value.

Below is a summary of the results after the KNN was applied and gender was inferred for the remaining customers in the data.

isMale_inferred_refined_model	Number_of_Customers	Percentage_of_base
0	35673	0.7749946
1	10357	0.2250054

It is important to note that the inferred gender from the Refined Model differs from those of the Baseline Model by only 4.4%. This is good news because the Baseline Model should already have been a good start, so a redical change by the Refined Model would have raised questions. The fact that the male/female split has not changed much either is also a positive sign. These figures could suggest that the Refined Model is indeed a “refinement”.

8. Prediction Model Trained on the Inferred Gender

Now that the gender has been inferred for the whole dataset, we can use this label to train a Prediction Model for future/other customers. KNN itself would be one option here again, but to display a variety of techniques for the purpose of this exercise, it was chosen to try out:

- a Logistic Regression Model,
- SVM models with different kernels, and
- a Neural Network

Little to no time was spent on fine-tuning the parameters of each algorithm, simply because the time was limited.

9. Prediction Model Evaluation

The models were trained on a random sample of 90% of the whole data set, then tested and evaluated on the remaining 10%. The performance of each predictive model was assessed on the Area Under the Curve (AUC) of both the Receiver-Operating-Characteristic (ROC) curve and the Precision-Recall (PR) curve, as well as the Accuracy.

Attention was given particularly to the AUC of the PR curve because the data set is not balanced. Indeed, suppose that the predictions were all “Female”. In that case the Accuracy would be around 77%, which by itself looks quite fine. A similar conclusion would be drawn from the AUC of the ROC curve. However, the AUC of the PR curve should highlight that such predictions perform far from fine.

Algorithm	aucROC	aucPR
LogicsticRegression	0.996849729255008	0.9908521759
SVM_poly_degree2	0.995495607503088	0.126503384157848
SVM_poly_degree3	0.991216328819594	0.126684357728257
SVM_radial	0.996867188140655	0.126475322847098
NeuralNet	0.999297147588425	0.997718299662934

Taking into consideration the results above and the efficiency of the algorithm, it was deemed that the best performing model was the Neural Network, even though each algorithm was highly accurate. Below is the summary of the Neural Network Model performance.

```
## Confusion Matrix and Statistics
##
##           Reference
## Prediction    0    1
##          0 3486    6
##          1   47 1064
##
##                Accuracy : 0.9885
##                  95% CI : (0.985, 0.9914)
##     No Information Rate : 0.7675
##     P-Value [Acc > NIR] : < 2.2e-16
##
##                   Kappa : 0.9682
##  Mcnemar's Test P-Value : 3.92e-08
##
##             Sensitivity : 0.9867
##             Specificity : 0.9944
##          Pos Pred Value : 0.9983
##          Neg Pred Value : 0.9577
##              Prevalence : 0.7675
##          Detection Rate : 0.7573
##    Detection Prevalence : 0.7586
##       Balanced Accuracy : 0.9905
##
##        'Positive' Class : 0
##
```

10. Summary

- The final outcome from this project is a Neural Network which predicts the gender of a customer, with an Accuracy of 98.8% over the inferred gender
- The Neural Network was trained on a set of customers with the inferred gender
- The methodology to infer the gender was based on a combination of heuristics and KNN
- Throughout the process, frequent checks against a baseline model were made to ensure consistency and relevance

Recommendation for additional features

1. Email address - to possibly extract customer name
2. Revenue split by item type - to know the distribution of revenue across item genders
3. Number of saved items in “Wishlist”, split by type - to have an idea of the customer’s potential future purchases
4. Total time on website - research suggests that females spend more time on site
5. Page visits per session - research suggests that females browse through more items

Suggested improvements for model performance

Here are just a few ideas on how the results could be improved further:

- Further investigation is needed on the data cleaning part, to have complete reconciliations
- The NAs need to be properly dealt with
- Including more features, such as the 5 suggested above
- Apply proper feature selection techniques; searching for multi-collinearity and possibly applying PCA for added insights
- Consulting industry experts to improve criterias for the *Core Labelled Subset*
- There is definitely room for improvement by applying cross validation and tweeking the model parameters
- Take a look at the coefficients of the Logistic Regression model - this could give some added insights on feature importance