RESELLER IDENTIFICATION FOR ONLINE SHOPPING BUSINESS

Zexi Han
Rue Data Science
Rue Gilt Groupe
Boston, MA 02210
zexihan@outlook.com

Lei Zhang
Rue Data Science
Rue Gilt Groupe
Boston, MA 02210
lzhang@ruelala.com

June 20, 2018

ABSTRACT

In this paper, we estimate the buyers' probability of being a reseller with feature engineering and machine learning from an iterative perspective.

***K*eywords** Feature Engineering · Machine Learning

1 Introduction

Our definition to resellers: A reseller is a member who purchases a great quantity of a particular style or a variety of styles and sell them for profits. (e.g. 6 same-style shoes of different sizes without return). They are valuable members but skew our personalization algorithms.

Basically, this is a binary classification task, a member is either a reseller or a normal buyer. Unlike phase 1 using rules for filtering, we take the machine learning approach to this problem in phase 2. The difficulties for this approach are listed below:

- Cold start: we don't know what features can best differentiate resellers from non-resellers, though Peter and Mehrnoosh's work in phase 1 gave us some inspirations on the features.
- No off-the-shelf dataset: we have to construct one.
- No clear boundary between resellers and active loyal buyers: in some cases, their purchase behaviors could be very similar, which also is a barrier when we do UAT to collect more data. It may take us a long time to look into the shopping history in order to determine whether the member is a reseller or not.

2 Iterative Machine Learning

To overcome these difficulties, we take an iterative machine learning workflow. First, we do feature engineering thinking and crafting representative features. Then if we have UAT data in the last iteration, we merge them to the dataset and remove them from the population data which is our 2 million buyers. After data preparation, we start modeling fitting a model with the best set of hyerparameters and evaluate its performance on the validation set. Eventually we use this model to predict the population data, and then we sample some members within certain probability bucket to do UAT. Based on the result of UAT, we got inspiration on the features. By putting the false positives and false negatives from UAT back to the dataset, we also got better data and more clear insight on the boundaries to start the next iteration of this workflow.

3 Patterns of Resellers

Before crafting features, we have to learn about the pattern of resellers. Based on my experience in UAT hundreds of members, I found there are basically two typical patterns or types of resellers. The first type of resellers concentrates on

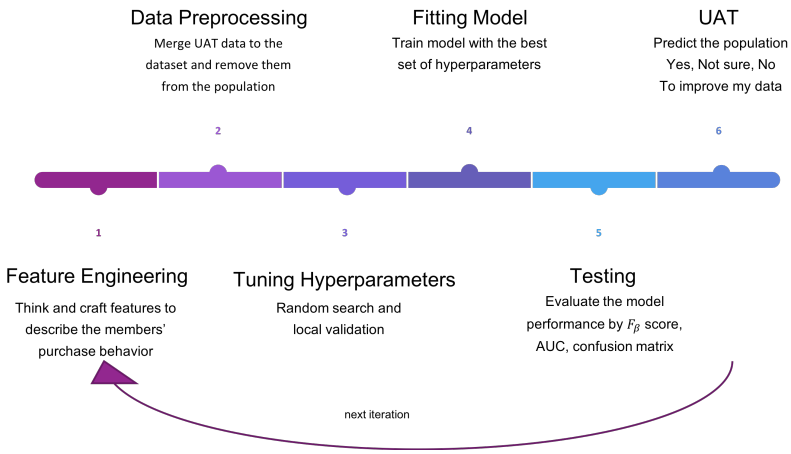

Figure 1: Iterative Machine Learning.

2/16/09	1111507564	Michael Kors MKS559 Black Rectangle Sungla..	Accessories	Sunglasses & Ey..	Sunglasses	Black	NS	$75	1	$0	0
	1111507566	Michael Kors MKS559 Tortoise Rectangle Sun..	Accessories	Sunglasses & Ey..	Sunglasses	Brown	NS	$75	1	$0	0
	1111515100	Michael Kors MKS580 Black Sunglasses	Accessories	Sunglasses & Ey..	Sunglasses	Black	NS	$75	1	$0	0
	1111515102	Michael Kors MKS580 Tortoise Sunglasses	Accessories	Sunglasses & Ey..	Sunglasses	Multi	NS	$75	1	$0	0
2/18/09	1111515309	Cole Haan Village Steeple Grey Large Cosmeti..	Health & Beauty	Make Up	Cosmetic Bags	Grey	Cosmeti..	$76	2	$0	0
	1111515357	Cole Haan Men's Black Leather ID Wallet	Accessories	Wallets	Wallets	Black	NS	$104	2	$0	0
	1311515427	Cole Haan "Air Astra" Ivory/Silver Patent Slip..	Shoes	Flats	Flats	Multi	11	$78	1	$0	0
2/23/09	1111513118	Kooba Black Leather Large Wallet with Studs	Bags & Wallets	Wallets	Wallets	Black	NS	$118	1	$0	0
	1111513169	Kooba Metallic Leopard Turnlock Wallet with ..	Bags & Wallets	Wallets	Wallets	No Color	NS	$236	2	$0	0
	1111513183	Kooba "Josephine" Black Leather Binded Clut..	Handbags & W..	Handbags	Clutches	Black	NS	$198	1	$0	0
	1111513184	Kooba "Josephine" Camel Leather Binded Clu..	Handbags & W..	Handbags	Clutches	Camel	NS	$198	1	$0	0
3/12/09	1111517693	Via Spiga "Furato" Tangerine Leather Hobo	Handbags & W..	Handbags	Hobo	Orange	hobo	$168	1	$0	0
	1111517721	Via Spiga "Lustra" Solar Patent Leather Large..	Handbags & W..	Handbags	Hobo	Yellow	hobo	$168	1	$0	0
3/26/09	1111513219	Kooba "Bailey" Silver Patent Triangle Frame ..	Handbags & W..	Handbags	Shoulder	Silver	shoulde..	$129	1	$0	0
	1111513220	Kooba "Bailey" Ruby Patent Triangle Frame C..	Handbags & W..	Handbags	Shoulder	Red	shoulde..	$258	2	$0	0
	1511499652	Spyder Girl's "Serpentine" Red Jacket	Girls Clothing	Tops	Tops	Multi	16	$29	1	$0	0
	1511499656	Spyder Girl's "Lightning" Red Jacket	Girls Clothing	Tops	Tops	Multi	20	$39	1	$0	0
	1511507370	Spyder Girl's "Force" Red & White Snow Pant	Girls Clothing	Bottoms	Bottoms	Multi	20	$39	1	$0	0
4/10/09	1111523155	Isabella Fiore "Let It Zip" Emana Hobo	Handbags & W..	Handbags	Hobo	Pewter	NS	$199	1	$0	0
	1111523215	Isabella Fiore "Rich Stitch" Celine Satchel	Handbags & W..	Handbags	Satchels	Copper	NS	$199	1	$199	1
4/14/09	1111521864	Gustto "Pavia" Green Leather East/West Satc..	Handbags & W..	Handbags	Satchels	Green	satchel	$249	1	$0	0
	1111521931	Gustto "Baca" Maize Yellow Leather Satchel	Handbags & W..	Handbags	Satchels	Yellow	satchel	$249	1	$0	0

Figure 2: Reseller Pattern 1.

only a few particular classes and they buy multiple sizes or colors for each style without return. In the screenshot, it is obvious that this member purchased multiple sizes for shoes. Why would a member purchase shoes of sizes from 8 to 11? Definitely for selling. I like this type of resellers because it is very easy to be identified.

The second type of resellers are not that friendly as the former one. They purchase a variety and a large quantity of products, probably 1 sku for each style. Look at this screenshot. This member totally purchased over 8 thousand 4 hundred items and spent over 1 hundred thousand dollars. And you will find that these orders are from a variety of categories and the quantity for each order is usually 1 or 2. But if we look more carefully, we can find that this member made too many purchases of various handbags. We may speculate that this member is a reseller.

Following the second type, the member below totally purchased over 2 thousand 7 hundred items and also spent over one hundred thousand dollars. The behavior of this member is confusing. Though it is similar to the last one, we cannot find any clue to identify this member as a reseller. For such cases, it is hard to tell whether he/she is a rich and loyal member or a reseller.

4 Feature Engineering

According to above patterns, we crafted 126 features and grouped them into 6 categories, which are general info, hierarchy percentiles, style and sku, site activities, geographical info and department buckets. The feature importance plot is shown below:

6/16/12	1414843465	Dolce & Gabbana Multicolor Logo Print Scarf	Clothing	Swimwear	Swimwear	Multi	NS	$49	1	$0	0
	6020843177	Links of London Silver Trumpet Charm	Jewelry	Necklaces	Necklaces	No Color	NS	$40	1	$0	0
6/19/12	1111851741	Chanel Luxury Ligne Bone Tote	Handbags & W..	Handbags	Handbags	Cream	NS	$2,299	1	$0	0
6/27/12	1311845479	Isola "Damani" Leather Sandal	Shoes	Sandals	Heels	Silver	6	$69	1	$0	0
	1313754916	Casadei Leather Sandal	Shoes	Sandals	Heels	Leopard	36.5	$260	1	$0	0
6/28/12	1411867613	Mikael Aghal "Fit 'n Flare" Silver Sequined Dr..	Clothing	Dresses	Occasion	Silver	8	$150	1	$0	0
	3010858156	Cuisinart Stainless Steel Citrus Juicer	Kitchen	Small Appliances	Small Applia..	No Color	NS	$30	1	$0	0
	3040775302	Ben's Garden "If You Live To Be A Hundred" H..	Gifts	Gifts & Collectib..	Gifts & Collec..	No Color	No Size	$80	1	$0	0
7/8/12	1111806585	FENDI "Mia" Denim Chain Handle Tote	Handbags & W..	Handbags	Totes	Blue	NS	$899	1	$0	0
7/14/12	1411763061	Josie Natori Periwinkle Matte Jersey Dress	Clothing	Dresses	Day	Periwinkle	Medium	$150	1	$0	0
	1411871069	SPANX "Golden Touch" One-Piece	Clothing	Swimwear	One-Pieces	Black	8	$80	1	$0	0
	6020827885	Majorica Silver Hoops	Jewelry	Earrings	Earrings	No Color	NS	$80	1	$0	0
7/19/12	1111861533	FENDI Women's FS5084 Sunglasses	Accessories	Sunglasses & Ey..	Sunglasses	Grey	NS	$150	1	$0	0
	1111878859	FENDI Women's FS5078 Sunglasses	Accessories	Sunglasses & Ey..	Sunglasses	Black	NS	$150	1	$0	0
	1411837215	7 For All Mankind "Kaylie" Dazzling Drake Bo..	Clothing	Jeans	Bootcut	Denim	30	$90	1	$0	0
7/20/12	3010828391	Brookstone Bluetooth Keyboard with Portfoli..	Gifts	Tech & Gadgets	Tech & Gadge..	No Color	NS	$140	2	$0	0
7/31/12	6020762641	Barbara Bixby 18K & Silver Gemstone Ring	Jewelry	Rings	Rings	No Color	7	$80	1	$0	0
	6020833507	Barbara Bixby 18K & Silver Pearl Initial Pendant (A-Z)	Jewelry	Necklaces	Necklaces	letter A	NS	$80	1	$0	0
						letter K	NS	$80	1	$0	0
8/4/12	1411860410	Hale Bob Brown Leopard Silk V-Neck Dress	Clothing	Dresses	Dresses	Multi	Medium	$130	1	$0	0
	1411884029	TART Collections "Bordeaux" Red Wrap Dress	Clothing	Dresses	Day	Red	Medium	$80	1	$0	0
	3010844490	Vera Wang for Wedgwood "With Love" 8in Va..	Gifts	Gifts & Collectib..	Gifts & Collec..	No Color	NS	$40	1	$0	0
8/5/12	3010807364	Mesa 19in French Loop Lazy Susan with Platter	Dining & Entert..	Serveware	China	No Color	NS	$52	2	$0	0
	3010848391	Brookstone Mobile Mini Speaker with Rubber ..	Gifts	Tech & Gadgets	Tech & Gadge..	No Color	NS	$32	1	$0	0
	3010885859	Brookstone "Surround Sound" Earbuds	Gifts	Tech & Gadgets	Tech & Gadge..	No Color	NS	$23	1	$0	0
8/6/12	4120741719	Corioliss "Baby SXE" Black Root Lifter	Health & Beauty	Tools & Accesso..	Tools & Acce..	No Color	NS	$20	1	$0	0
	4120781523	Corioliss Straight 2 Curl Black Styling Iron	Health & Beauty	Tools & Accesso..	Tools & Acce..	No Color	NS	$100	1	$0	0
	6010878221	Ferragamo Women's Gancino Watch	Watches	Straps	Leather	No Color	NS	$460	1	$0	0

Figure 3: Reseller Pattern 2.

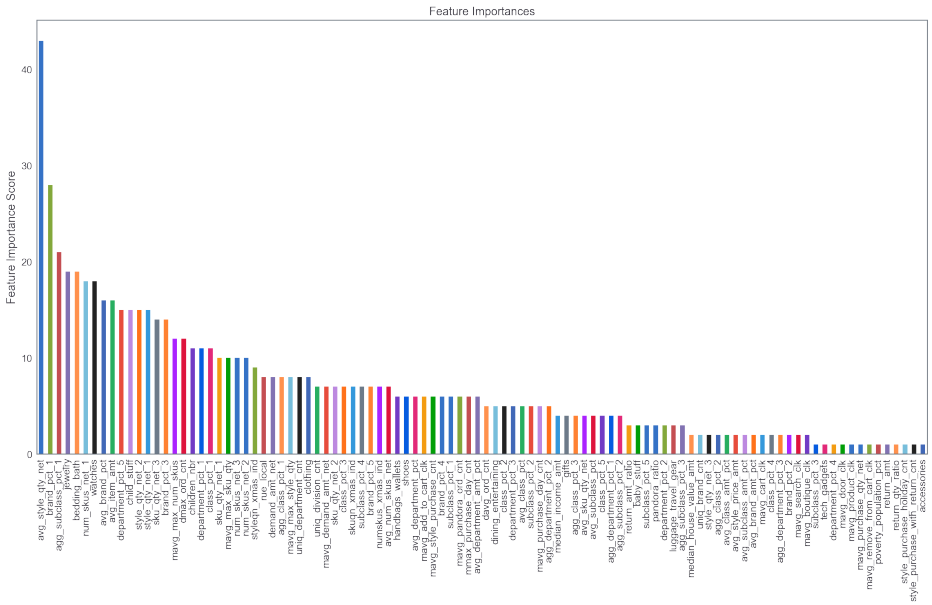

Figure 4: Feature Importance.

4.1 General Info

To represent the basic purchase activity, we aggregated some general statistics of purchases: The count of style level purchases. The count of style level purchases with return. The average price of styles The net quantity of purchased items. The quantity and amount of returned items. The ratio of returns. The average monthly quantity of above features. The max/average monthly count of days with purchases. Total amount of money spent. The max/average daily count of orders. The count of unique divisions / departments / classes / subclasses. To differentiate the confusing Pandora buyers from others, we also have the count and the ratio of Pandora orders. The number of children is helpful to differentiate members who purchased a lot of baby and children stuff.

4.2 Hierarchy Percentiles

Peter and Mehrnoosh's work in phase 1 gave me an inspiration that we can check the cumulative distribution of a style order quantity in a group of quantities under the same RDS hierarchy. For example, if there is one single order of 20 same style Pandora charms in a member's shopping history, while most of Pandora buyers usually buy 1 or 2 same style charms. Then this member is more likely to be a reseller than others. This idea creates the features of top 5 and average percentiles in department level. And we have such percentiles in department, class, subclass and brand level. The purpose of putting top 5 rather than a single max here is to reduce false positives. We can tolerate one or two suspect purchases. If a member only has 1 or 2 style level purchases of high percentiles while all other percentiles are low, he/she shouldn't be identified as a reseller. Besides, we also have the aggregated department percentile which is the cumulative distribution of the aggregated sum of order quantity grouped by RDS hierarchy in a group of quantities under the same RDS hierarchy. For example, if one totally purchased 500 items in clothing department, while most clothing buyer purchased 10 or 20 from clothing department, then this buyer is more likely to be a reseller. Of course, such member could also be an active buyer. So, this feature is weaker than the above.

4.3 Style and Sku

The style and sku features are important in identifying the first type of resellers. We have: Top 3 style level number of skus, sku quantity, style quantity Average number of skus, sku quantity, style quantity Average of max number of skus per month, max sku quantity, max style quantity The Christmas indicator feature indicates whether these top 3 orders happened in holiday season.

4.4 Site Activities

For site activities, we have the click count feature from login to checkout.

- Login click
- Search click
- Boutique click
- Door click
- Product click
- Add-to-cart click
- Remove-from-cart click
- Cart click
- Checkout click

4.5 Geo Info

In terms of geographical information, we map the zip code of members' primary order address to the census data and thus got three features. The median amount of income. The median house value. The poverty population percentile.

4.6 Department Buckets

We created 31 department buckets and sum the quantity of items for each bucket. Child stuff and baby stuff are aggregated buckets. By creating these features, we want to provide the information of what categories the members usually make purchases from. Because members who majorly make purchases from different category may have

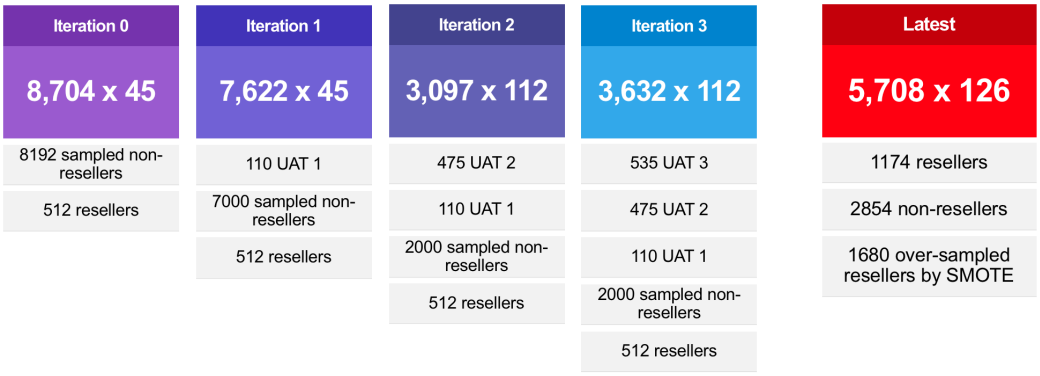

Figure 5: Evolution of the Dataset.

different behaviors. For example, some members usually buy multiple items for each style from baby stuff bucket while some others only buy one item for each style from handbags and wallets bucket.

5 Dataset Construction

Better data beats better algorithms. Since we don't have a dataset, we have to create one on our own.

5.1 Initiation

Initially, with Peter's algorithm, we found 512 members that make purchases of a magnitude and we assume that they are just resellers. To find non-resellers, we relaxed certain cut-off value in Peter's algorithm to leave some space for resellers, and then randomly sampled 8 thousand members from 2 million buyers. We also assume that they are just non-resellers. The reason why we sampled 8 thousand rather than 5 hundred to make classes balanced is that we have to take advantage of this imbalance when we don't have many examples in dataset. In reality, the quantity of non-resellers is just much larger than resellers. More importantly, the variance in non-resellers is also much larger than resellers. It means we should put more weights on non-resellers so as to predict the correct probability. And according to my experiments, 1 to 16 is a proper ratio to sample non-resellers and as a result we can get around 3 thousand suspect resellers out of 2 million buyers with probability larger than 0.5 in our first model.

In the following iterations, we put the false positives and true positives in UAT back to the dataset. Meanwhile, we replaced a large number of sampled non-resellers with a small number of false positives in UAT. In this way, we can put more weights to the false positives in the dataset in order to reduce the quantity of false positives. Because for this binary classification problem, we prefer miss resellers rather than misclassify non-resellers as reseller.

In the latest dataset, we have 5708 examples, 11 hundred resellers and 28 hundred non-resellers. We over-sampled the minority class (the resellers class) by SMOTE (https://arxiv.org/abs/1106.1813), which is an over-sampling technique creating synthetic minority class examples rather than replacement, and is usually used in imbalanced classification problem. In this way, we can strengthen the features of resellers and better evaluate our model.

5.2 Find Non-Resellers

We used an adjusted version of Peter's distribution algorithm. The members who are identified as resellers and non-resellers will then be served as labeled data for machine learning approach to identify resellers.

Finding non-resellers we need is not a simple reverse of finding resellers. The number of non-resellers is much larger than resellers. Suppose we have selected five hundred members who are highly likely to be resellers. If we want to construct a dataset on which we run machine learning algorithms, we may firstly think about having the dataset with balanced distribution between positive and negative instances. So we need five hundred non-resellers to be selected. We can do this by sampling from the set of members who are temporarily believed to be non-resellers. To find such set, we define some thresholds on the distribution for quantity purchased. Problem solved.

It is noteworthy that the dataset created for modeling, in general, should better have a balanced distribution between positive and negative. If not, i.e. there are more non-resellers than resellers in the dataset, accuracy is not a suitable indicator to choose cutoff. We use ROC curve to decide which cutoff value to choose in order to make a balance

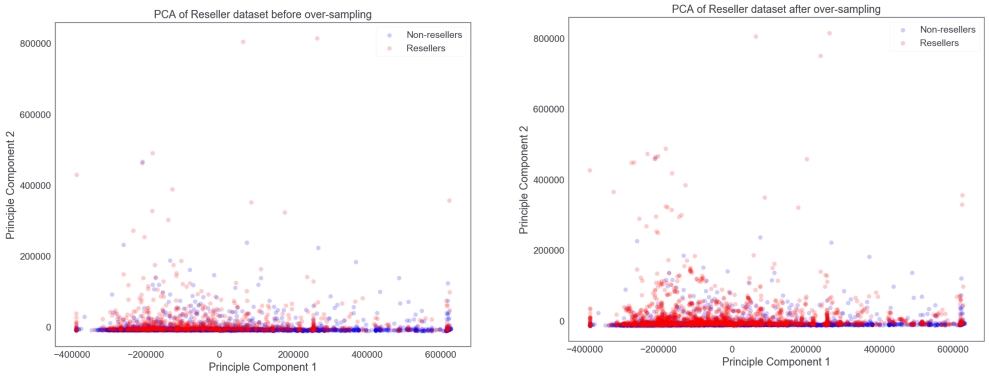

Figure 6: Principle Components of the Features Before and After Oversampling.

between the false positive rate (FPR) and false negative rate (FNR) when we have unbalanced distribution. Experiments are done on both cases.

6 Modeling

6.1 Exploration - Stacking

Initially, we don't know which model performs best in this problem. Depending on a variety of factors, different model families will perform better. Therefore, one of the easiest ways to improve our solution for a given problem is to try several different model families. We tried logistic regression, random forest, SVM, neural network and gradient boosting model. And each of them (except the gradient boosting, it has too many parameters) is tuned with cross-validation and grid search on hyperparameters.

The next way to improve the solution is by combining multiple models into an ensemble. Here we take the stacking by training a 2nd level logistic regression on the probabilities predicted by the model families. This is a direct extension from the model exploration. This combined prediction will often see a small performance increase over any of the individual model.

(Stacking uses a similar idea to k-folds cross validation to create out-of-sample predictions. The key word here is out-of-sample, since if we were to use predictions from the M models that are fit to all the training data, then the second level model will be biased towards the best of M models. This will be of no use.)

```
listOfFamilies = logistic regression,
                 random forest,
                 svm,
                 neural network,
                 xgboost

for model in listOfFamilies:
    bestModel = tuned with cross-validation on trainingData

stacking of models from modelFamily with a 2nd level logistic regression
```

6.2 Diving into XGBoost

Based on the model exploration in the first two iterations and the probability prediction result by the individual model and the stacking model, we found the performance of the gradient boosting model with the default setting is almost as good as the stacking model. So, we decided to focus on tuning one model instead of five. As we know, there are a bunch of hyperparameters in gradient boosting decision tree model. In terms of hyperparameters tuning, we tried traditional grid search and cross validation. But it would take hours to find a best set of parameters in my laptop and we cannot afford that. So we take a different approach - random search and local validation. We set the grid search points manually and random select one point from the set for each iteration. After training for like 1000 iterations,

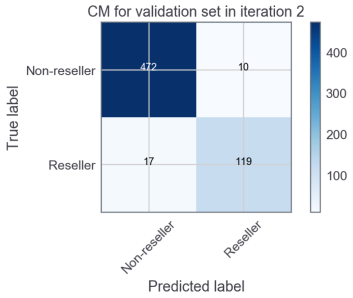

Figure 7: Confusion Matrix of the Inference Result on the Validation Set.

which would take 30 minutes, we observe the average performance of each candidate value for each parameter, make adjustments to the search points and repeat the training.

Different parameters control different function of the model. We have number of boosted trees, learning rate to control the learning progress, several others to control overfitting from different perspective, and scale positive weight to overcome the imbalance of data.

```
model = XGBoost

bestHyperparameters = tuned with random search and local validation

bestModel = model.fit(trainingData, bestHyperparameters)
```

7 Results and Discussion

7.1 Evaluation Metric

Since the imbalance nature of data and we value precision more than recall, we use F-beta score as our evaluation metric. It is a harmonic mean of precision and recall which lends more weight to precision when beta in a range 0 to 1. Actually, there are 3 influential factors to the weight of precision and recall – the imbalance of classes, scale-positive-weight parameter in model and this beta. The confusion matrix on the right shows the model performance on validation set in iteration 2. And we can see that the 10 false positives are less than the 17 false negatives.

$$F_\beta = (1 + \beta^2) \cdot \frac{precision \cdot recall}{(\beta^2 \cdot precision) + recall}$$

7.2 Model Performance

Since there were few examples in the dataset in early iterations, we split the dataset into training and validation set in a ratio of 8:2. Train model on the training set, tune parameters on the validation set and evaluate the performance by UAT. The drawback of this approach is that the actual model performance could not be obtained once the model was trained because we had to spend a week on UAT. Th bar chart below shows the F-0.5 score on training set and validation set from iteration 0 to iteration 3. We can see the F-0.5 score on the validation set was decreasing. It is normal because the dataset only contains conservatively selected 5 hundred resellers and randomly sampled 8 thousand non-resellers at the begining, so the model performance could be high. And with the project going on, we added more difficult data to confuse the model. We can also see that the model overfits the training set. To see the real performance of model, we shouldn't look at it on the validation set that is used for tuning hyperparameters. We should look at the performance on the test set which is UAT in our case.

In our latest version, there are much more examples than before so we split the dataset into training / validation / test set in a ratio of 70

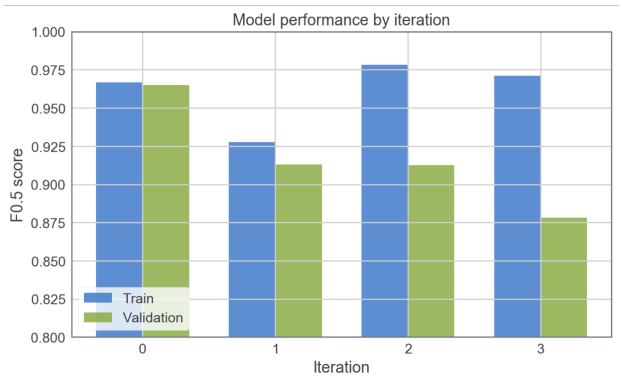

Figure 8: F0.5 Score Performance by Iteration.

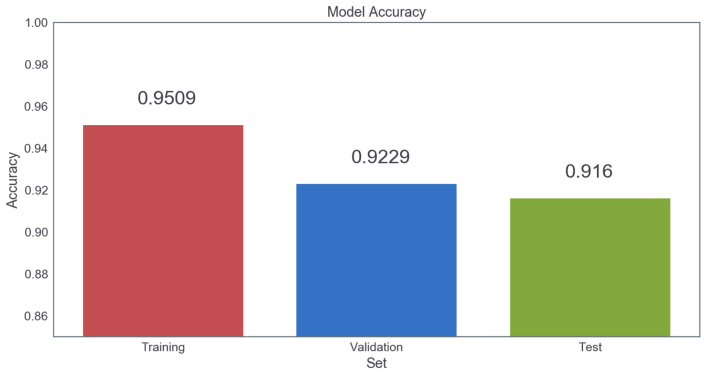

Figure 9: Model Accuracy Performance.

7.3 UAT Performance

In each iteration of UAT, we bucket the probability to select the members for UAT and check its accuracy performance in each UAT. We can clearly see that the improvement in accuracy from 0.5 to almost 0.9 in 0.9-1.0 bucket from the iteration 0 to iteration 4, which means the false positives are being reduced and also proves that adding more UAT data works. However, the accuracy in 0.4-0.5 bucket decreased a bit. But I believe this problem could be solved by adding more data to get a better generalization for model. Thanks to Parker, we have two tableau views (order date view and hierarchy view) to do UAT: https://us-east-1.online.tableau.com//site/ruelala/workbooks/234118/views.

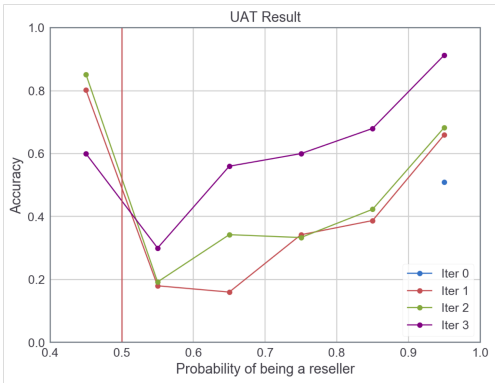

Figure 10: UAT Result by Iteration.

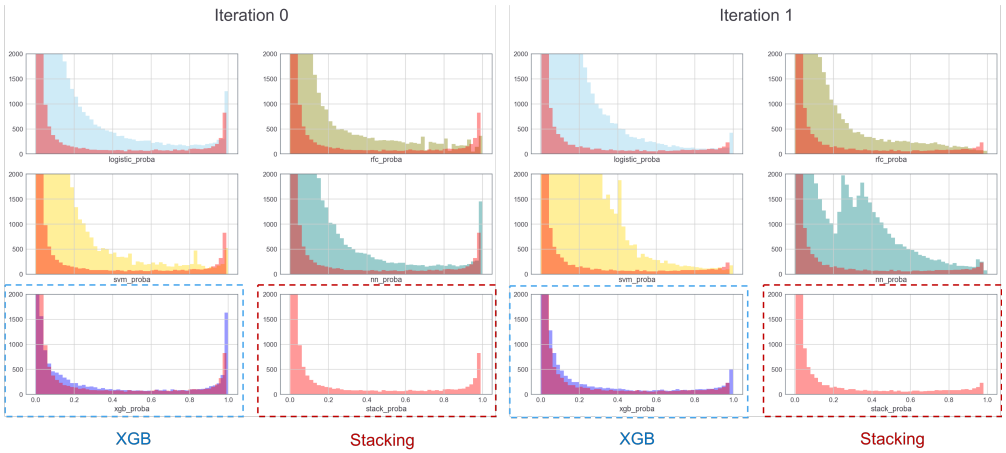

Figure 11: Model Selection.

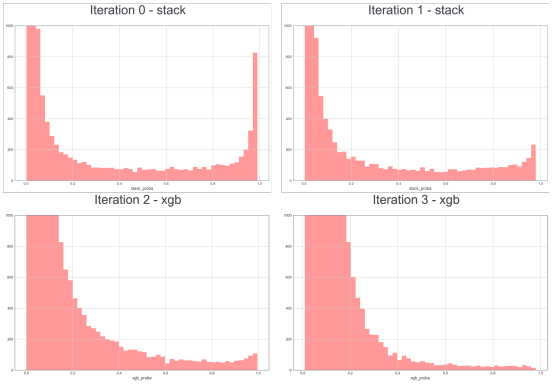

Figure 12: Reduction of False Positives.

7.4 Probability Distribution of the Predictions

7.4.1 Model Selection

We can also see this iterative improvement by checking the probability distribution of the predictions over our 2 million buyers. In the model exploration phase iteration 0, there is an obvious spike shape at the position of high probability. Based on the result of first iteration of UAT, we can say that there are a lot of false positives in the spike. After adding the 1 hundred UAT data to the dataset, the spike or the quantity of false positives shrinks lot. We also found that the single gradient boosting model performs almost as good as the stacking of five models.

7.4.2 Reduce False Positives

In this group of graphs, we can clearly see the drop of false positives. The quantity of members with probability above 0.5 drops from over 3 thousand to less than 1 thousand. We are confident that most of members in this drop are false positives rather than true positives because we have seen the rise in accuracy in both 0.4-0.5 and 0.9-1.0 buckets in UAT.

8 Deployment

Training Module

Frequency: Per month

When training module is called, it will do the following tasks automatically:

- Request training data from Snowflake

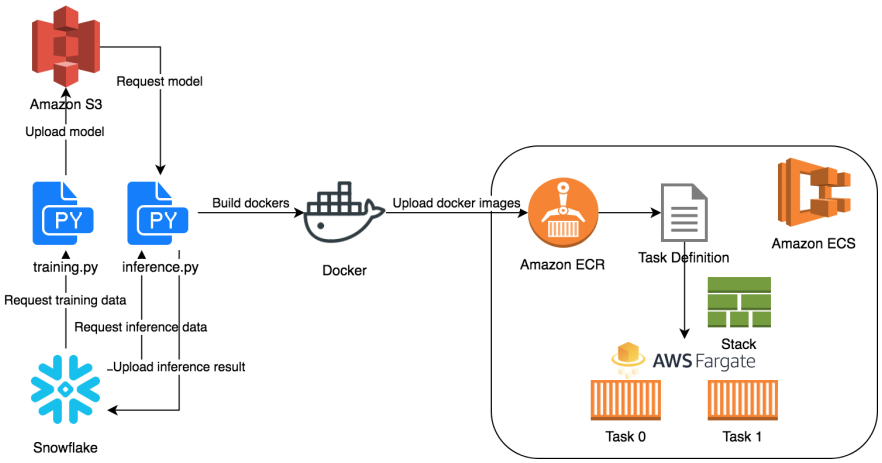

Figure 13: AWS Workflow.

- Split training/validation/test set
- Over-sampling
- Tuning parameters
- Fit best model
- Evaluate model performance
- Upload the model and log file to S3
-

Log file is saved to S3 so as to check the training process manually.

Inference Module

Frequency: Per week

When inference module is called, it will do the following tasks automatically:

- Request inference data from Snowflake
- Request model from S3
- Predict and estimate the probability
- Upload the inference result to Snowflake

Log file is saved to S3 so as to check the inference process manually.

Appendix - Features

General Info

General statistics describing members' purchasing behavior.

- 'is_reseller': binary label
- 'style_purchase_cnt': the count of purchased distinct styles
- 'avg_style_price_amt': the average demand amount for each purchased style
- 'style_purchase_with_return_cnt': the count of purchased distinct styles with return
- 'style_purchase_holiday_cnt': the count of holiday (Nov and Dec) purchased distinct styles
- 'purchased_qty_net': the net quantity of purchased items
- 'returned_qty': the quantity of returned items

- 'return_amt': the amount of returned items
- 'return_amt_ratio': the ratio of the amount returned to the amount ordered
- 'avg_item_amt': the average amount of each purchased sku
- 'mavg_style_purchase_cnt': average monthly count of purchased distinct styles
- 'mavg_purchased_qty_net': average monthly net quantity of purchased items
- 'mavg_returned_qty': average monthly quantity of returned items
- 'returned_ratio': the ratio of returns - returned quantity divided by total quantity
- 'mmax_purchase_day_cnt': max monthly count of days with purchases
- 'mavg_purchase_day_cnt': average monthly count of days with purchases
- 'mavg_price_amt': average monthly money spent
- 'pandora_ord_cnt': the count of pandora orders
- 'pandora_ratio': the ratio of pandora orders to all orders
- 'dmax_ord_cnt': max daily count of orders
- 'davg_ord_cnt': average daily count of orders
- 'uniq_division_cnt': the count of unique divisions
- 'uniq_department_cnt': the count of unique departments
- 'uniq_class_cnt': the count of unique classes
- 'uniq_subclass_cnt': the count of unique subclasses
- 'children_nbr': the number of children of the member

Hierarchy Percentiles

Top 5 style purchase level: the cumulative distribution of a style order quantity/amount in a group of quantities under the same RDS hierarchy.

- 'avg_department_pct'
- 'avg_class_pct'
- 'avg_subclass_pct'
- 'avg_brand_pct'
- 'department_pct_1'
- 'class_pct_1'
- 'subclass_pct_1'
- 'brand_pct_1'
- 'department_pct_2'
- 'class_pct_2'
- 'subclass_pct_2'
- 'brand_pct_2'
- 'department_pct_3'
- 'class_pct_3'
- 'subclass_pct_3'
- 'brand_pct_3'
- 'department_pct_4'
- 'class_pct_4'
- 'subclass_pct_4'
- 'brand_pct_4'
- 'department_pct_5'

- 'class_pct_5'
- 'subclass_pct_5'
- 'brand_pct_5'
- 'avg_department_amt_pct'
- 'avg_class_amt_pct'
- 'avg_subclass_amt_pct'
- 'avg_brand_amt_pct'

Top 3 hierarchy purchase level: the cumulative distribution of the aggregated sum of order quantity grouped by RDS hierarchy in a group of quantities under the same RDS hierarchy.

- 'agg_department_pct_1'
- 'agg_department_pct_2'
- 'agg_department_pct_3'
- 'agg_class_pct_1'
- 'agg_class_pct_2'
- 'agg_class_pct_3'
- 'agg_subclass_pct_1'
- 'agg_subclass_pct_2'
- 'agg_subclass_pct_3'

Style and Sku

Top 3 style level number of skus.

- 'num_skus_net_1'
- 'num_skus_net_2'
- 'num_skus_net_3'
- 'num_skus_xmas_ind'
- 'avg_num_skus_net'
- 'mavg_max_num_skus'

Top 3 sku level sku quantity.

- 'sku_qty_net_1'
- 'sku_qty_net_2'
- 'sku_qty_net_3'
- 'sku_qty_xmas_ind'
- 'avg_sku_qty_net'
- 'mavg_max_sku_qty'

Top 3 style level style quantity.

- 'style_qty_net_1'
- 'style_qty_net_2'
- 'style_qty_net_3'
- 'style_qty_xmas_ind'
- 'avg_style_qty_net'
- 'mavg_max_style_qty'

Site Activities

The number of clicks on different area of the web front.

- 'mavg_login_clk'
- 'mavg_search_clk'
- 'mavg_boutique_clk'
- 'mavg_door_clk'
- 'mavg_product_clk'
- 'mavg_add_to_cart_clk'
- 'mavg_remove_from_cart_clk'
- 'mavg_cart_clk'
- 'mavg_checkout_clk'

Geo Info

Simple geographical information regarding members' purchasing capacity.

- 'median_geo_income_amt'
- 'median_geo_house_value_amt'
- 'geo_poverty_population_pct'

Department Buckets

Style level quantity in 31 department buckets.

- 'accessories'
- 'handbags_wallets'
- 'luggage_travel_gear'
- 'bags_wallets'
- 'tech_gadgets'
- 'health_beauty'
- 'shoes'
- 'grooming_cologne'
- 'bedding_bath'
- 'clothing'
- 'dining_entertaining'
- 'home_decor'
- 'gifts'
- 'holiday'
- 'kitchen'
- 'outdoor'
- 'pets'
- 'loyalty'
- 'programs'
- 'travel'
- 'furniture'
- 'jewelry'

- 'watches'
- 'sports'
- 'costumes'
- 'books'
- 'rue_local'
- 'wine'
- 'storage_cleaning'
- 'child_stuff': aggregated bucket
- 'baby_stuff': aggregated bucket