INTRODUCTION

Mail marketing campaign are often used to target existing as well as potential new customers with aim to increase profits in the short but also in the long run by investing in customer loyalty and continues communication with potential customers. XYZ company, who has been conducting mail marketing campaign in the past years, is looking to deploy a 17th mail campaign and is interested in learning if they should be using a different and new targeting approach that will yield better profit for the 17th campaign than they obtained through the 16th mail campaign. Together with exploring, processing and cleaning the customer data provided by XYZ as well as purchased from third party organizations, we are asked to build and assess supervised machine learning models to determine whether or not XYZ company need to change their targeting strategy or continue using the current methods.

EXPLORATORY DATA ANALYSIS

The dataset provided by the XYZ company consists of 30,779 observations and 554 variables. The dataset contains excessive number of 554, that is a combination of data collected by XYZ company on customer transaction data and demographic data on each customer that was purchased from Experian. The customers are geographically reside in one of the 36 unique zip codes in the dataset which are all located in the Chicago area. As we are looking to base our study on the last mail campaign performance, response16 variable is our response variable, and we will continue our exploratory data analysis (EDA) starting with responser16 and exploring variables that we think are related to it. The dataset is dated 2009, and hold the information on 15 years prior to 2009, making the mail campaign in 2009 the 16^{th} campaign, hence the response16 variable. As the number of exploratory variables is huge, we are aiming to use our domain knowledge to draw the clear path to data exploration and then use the key variables in our analytical models. We are also hoping that EDA gives us the overview on which of the relevant variables require additional cleaning and processing prior to logistic regression and random forest model conduction.

As the initial step of our EDA, we will be looking at variables that are related to the status of the last campaign and the customer. We see that from all 30,779 customers in our dataset 14,922 of them, which is 48% of the total, received mail through the latest campaign. When we look further into this, we see that out of 14,922 customers who received the mail only 1,440, which is 10%, responded. Interestingly, when we look at all the customers who responded, we see that there are an additional 1,152 customers who did respond even though they didn't receive

any mail during the latest campaign. This insight shows us that the response rate is reasonable similar between customer who did and did not receive any mail in the last campaign.

Another significant variable to explore related to customer is their status. In the dataset the status of a buyer is identified as either Active, Inactive or Lapsed. We see that 13,602 customers are identified as active, while 17,177 of them, about 56% of all customers, are either inactive or lapsed. When we further look into who within these buyer status groups received mail in the last campaign (any_mail_16) we see that the total email sent to active customers is almost as same as the mail sent to inactive and lapsed customers. In addition, only 56% of the active customers received mail in the last campaign while almost half of the active customers didn't. When the customer status information is explored in customer response context, all 2,592 responses received to the last mail campaign were from customers with active status.

	RECEIVED	
	0	1
ACTIVE	6046	7556
INACTIVE	5439	3722
LAPSED	4372	3644

	RESPONDED	
	0	1
ACTIVE	11010	2592
INACTIVE	9161	0
LAPSED	8016	0

Table 1: Recipient and Response by Buyer Status

A few of the key variables related to customer demographics that we explore prior to data processing and cleaning is Median Income, Age and their Home Ownership. Per Figure 1 below, the median income ranges mainly between $40,000 and $120,000 and the customers are mostly aged between 35 and 70. With either of the histogram plot we observe no extreme outliers and a strong skew as a result of it, yet we observe a possible minor outlier condition with median income which we will further explore.

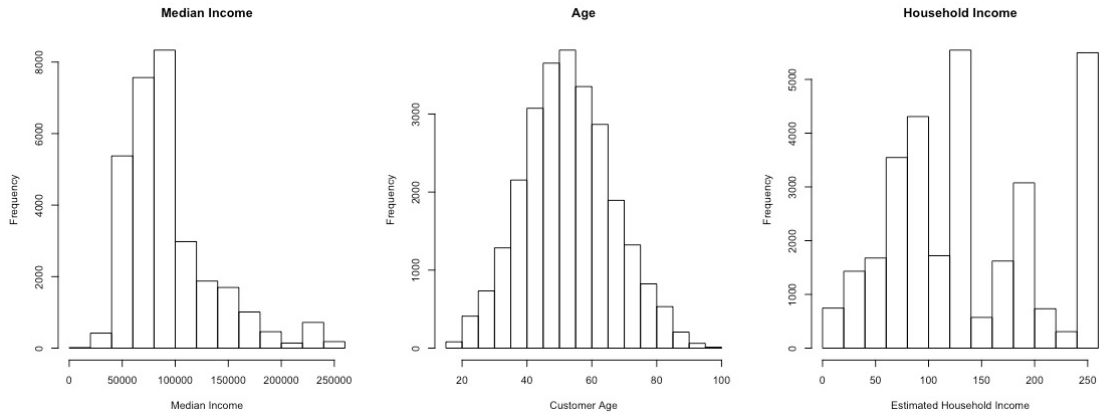

Figure 1: Histogram Plot for Estimated Household Income, Median Income and Age

Furthermore, when we investigate which customer age group have provided the most response to campaign 16, we see that customers age group of 45 to 60 has been the most responsive. Also, when the conduct the same investigation with median income we see that in campaign 16 most responses were received from customers with median income of $100,000 and above. Finally, the estimate household income plot indicates a rather scattered pattern with an accumulation observed around $120,000-$140,000 range and above $250,000.

Following our exploration of latest campaign mailing, response rate and customer status and demographics in relation to these variables, we look at the customer transaction and sales information. Looking at the historical sales and transaction data by customer dating pre 2009 we see that mean sale per customer is $979, while median and maximum sale per customer is $426 and $94,350, respectively. The extensive difference between the mean and maximum indicates a skew caused likely by an outlier. Additionally, if we look at the transaction count in order to see the bigger picture, we see that on average a customer makes 3.8 transactions, and similar to the summary statistics we obtained by looking at pre 2009 sales, the median and maximum transaction per customer is 2 and 178, respectively. Same skewness issue is observed with transaction information where the mean and maximum are too apart from each other which indicate outliers. In our attempt to visualize outliers we include the median income variable that we previously plotted which indicated possible outliers.

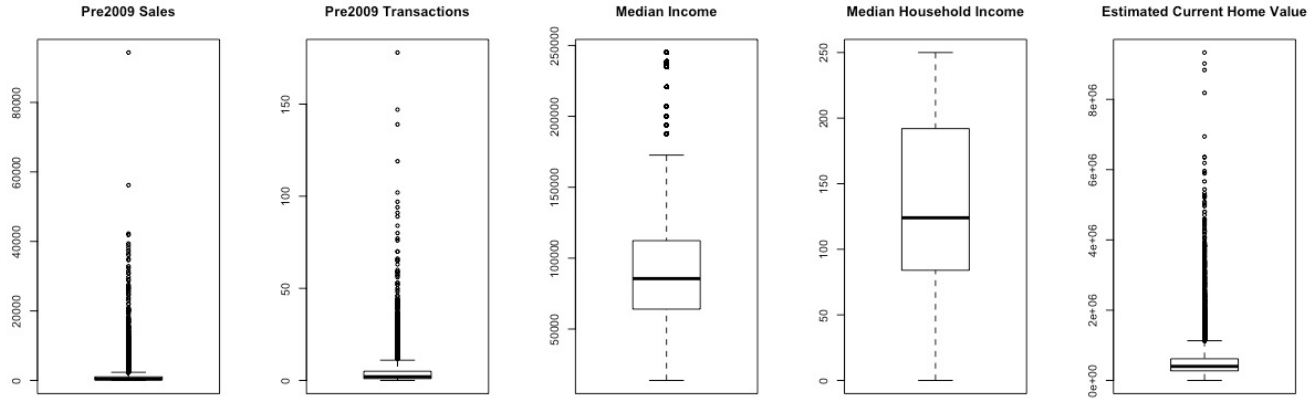

Figure 2: Boxplots for Pre2009 Sales, Transactions and Median Income

While Median Income and Estimated Household Income explanatory variables do not show any extreme outliers, we should be concerned about, the historical sales, transaction and estimated home value data do indicate several extreme outliers which need removal.

Per the insights obtained through exploratory data analysis and the cumulative industry knowledge we choose several explanatory variables for further processing and modelling. The variables mentioned are as follows: PRE2009SALES (Total of all sales per customer dating pre 2009), PRE2009TRANSACTIONS (Total number of transactions per customer dating pre 2009), MED_INC (Median Income), cum15QTY (Cumulative total amount), QTY15 (Quantity), cum15TOTAMT (Cumulative total amount of sale), TOTAMT15 (Total amount of sale), ESTHMVL (Estimated Home Value) , EXAGE (Age), INC_WOUTSCS_AMT_4 (Household Income), ZKITCHEN (Kitchen Aids/Small Appliances), SUM_MAIL_16, NUMBADLT (number of adults), TOTAL_MAIL_16 (Number of mail received in campaign 16), ANY_MAIL_16 (Campaign 16 mail recipience,RESPONSE16 (Campaign 16 response). Additionally, we are creating two new variables that are Sale per Campaign by dividing total sales with number of campaigns, and Sale per Transaction, that we obtain through dividing sales with the transaction count.

DATA CLEANING & PROCESSING

As the first step of our data cleaning, we remove all observations that have 0 for ANY_MAIL_19 variables, which indicates that they haven't received mail during the latest campaign. This step already turned the 30,779 observations dataset into a 14,922-observation set. Within the new dataset explanatory variables Estimated Home Value, Age, Sale per Campaign, and Sale per Transaction have missing data. As the missing data is likely to impact the performance of the logistic regression and random forest models we aim to conduct, a data imputation exercise is necessary. As none of these variables have missing data that is more than 20% of the total observations, which is 14,922, we move forward with data imputation by using the mean value.

Another variable that demands minor cleaning is ZKITCHEN. The variable is stored in the raw data as character type variable with values "U", which we assume to indicate unknown, and "Y", which is Yes. There also numerous observations with NAs. We transform the variable by changing "Y" into 1 and "U" and NAs into 0. With this exercise we now know that out of 14,922 observations 5012 of them have kitchen. While this condition is less than ideal, this variable is crucial for our model as the campaign handout exhibits large and small kitchen appliances. As the final attempt to recover this variable, we create a dummy variable called ZKITCHEN_U where what we assume NAs in variable ZKITCHEN will be moved and will be stored in binary format.

The data cleaning and processing exercise for the XYZ dataset prior to modelling involves cleaning outliers in variables Pre 2009 sales, transactions and estimated home value. As presented in Figure 2 above, we determined that there are only few extreme outliers in these variables that may skew the data. In order to achieve this we remove the variables that are outside of 1st and 99th percentile. With this exercise the boxplot diagrams presented in Figure 2 were changed to plots presented in Figure 3 below.

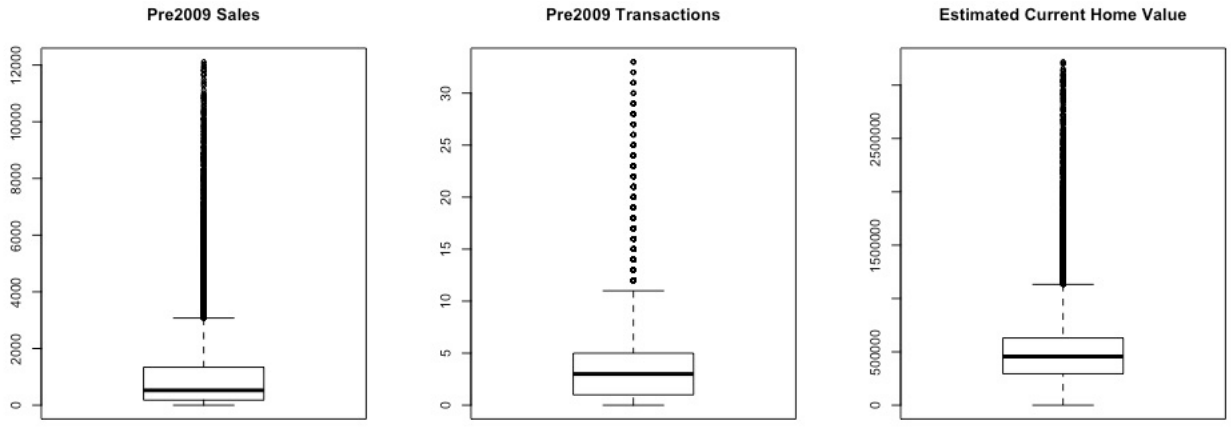

Figure 3: Boxplots for Pre2009 Sales, Transactions and Median Income Post Outlier Removal

Following the above transformations of data imputation, outlier removal and data type change, the dataset we will be using in our models has 19 variables and 14,695 observations.

In conclusion of the EDA, cleaning and processing of data, when we review the RESPONSE16 variable we see that out of 14,776 customers only 1,386 responded meaning app. 90% of the customers did not respond. This still indicates a highly skewed dataset yet due to low response rate which is the nature of the collected data.

MODELLING

Following the above transformations of data imputation, outlier removal and data type change, the dataset we use in our models consists of 14,695 observations and 19 variables. Our goal is to run two models, a logistic regression model and a random forest model where in both we will model the response variable RESPONSE16 with PRE2009Sales, PRE2009Transactions, MED_INC, EXAGE, TOTAMT15, cum15QTY and salepertransaction. The variables are selected per their relevance to both customer sales and transaction history, campaign history and demographics.

The generalized linear model with binominal regression that was conducted with the response and explanatory variables listed above outputs null deviance of 9181.9 on 14,694 degrees of freedom and residual deviance of 8517.7 on 14,687 degrees of freedom and an AIC score of 8533.7. The residual deviance is 664 less than the null deviance, suggesting that the explanatory variables included in the model are good predictors. This is also confirmed by the low p-values of the majority of the variables included in the model which indicate significance. The Hosmer and Lemeshow Goodness of Fit (GOF) test outputs x-squared value of 50.35 and a p-value less than 0.05. These outputs suggest a poor fit for our logistic regression model. Additionally, the logistic regression model is highly accurate in predicting 'no response' yet fails to predict customers who are likely to respond to the mail campaign. The model has 90% accuracy in predicting the customers who are not likely to respond, yet when we review the confusion matrix, we see that it modelled over 96% of the customers who have responded to campaign 16 as customers who are not likely to respond. Figure 4 exhibits the histogram plot of logistic regression model prediction accuracies.

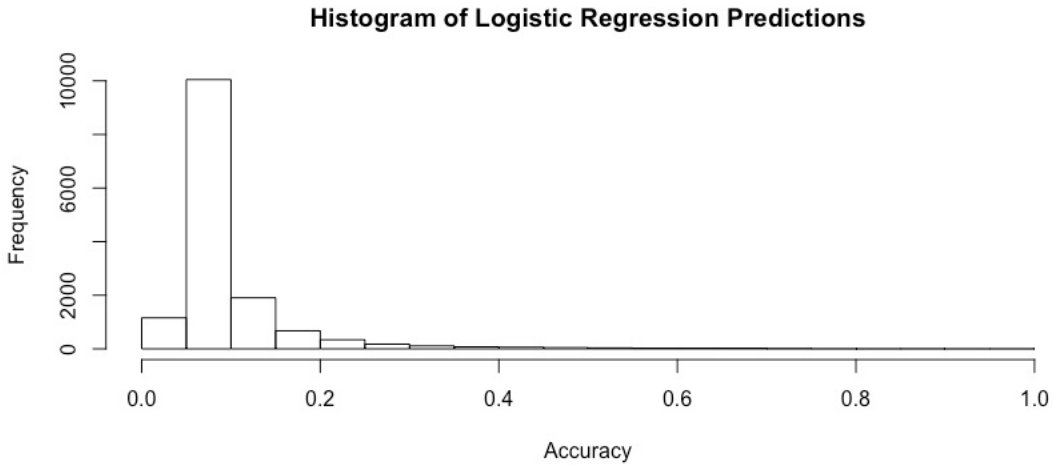

Figure 4: Histogram of Logistic Regression Prediction Accuracies

Per the histogram plot we see that the most frequently seen prediction accuracies are between 0% and 20% and there are very few prediction accuracies above 50%. This plot not only shows skewed results but also an inaccurate model in predicting the response likelihood of a customer.

We conduct a second model, Random Forest (RF) model which will use numerous decision trees to identify variables that are significant in predicting RESPONSE16. For the RF model we use the same variables that was

used in the logistic regression model. As the initial step we examine the confusion matrix generated by looking at RESPONSE16 against Random Forest Model predictions.

		RF Predictions	
		0	1
RESPONSE16	0	13309	0
	1	539	847

Table 2: Confusion Matrix of Response16 vs. Random Forest Predictions

Per Table 2, we see that random forest models shows a perfect accuracy of predicting 'no responses' and a better performance in regard to predicting the likeliness of customer response to the mail campaign. While there are still almost 40% of the predictions for response=1 misplaced under prediction=0, the matrix indicates that there are more observations with high prediction accuracy with random forest model than the logistic regression. The histogram plot below in Figure 5 also confirms this interpretation. We see that while the prediction accuracies are clustered between 0% and 20% and are highly skewed, there are prediction accuracies that are above 50%.

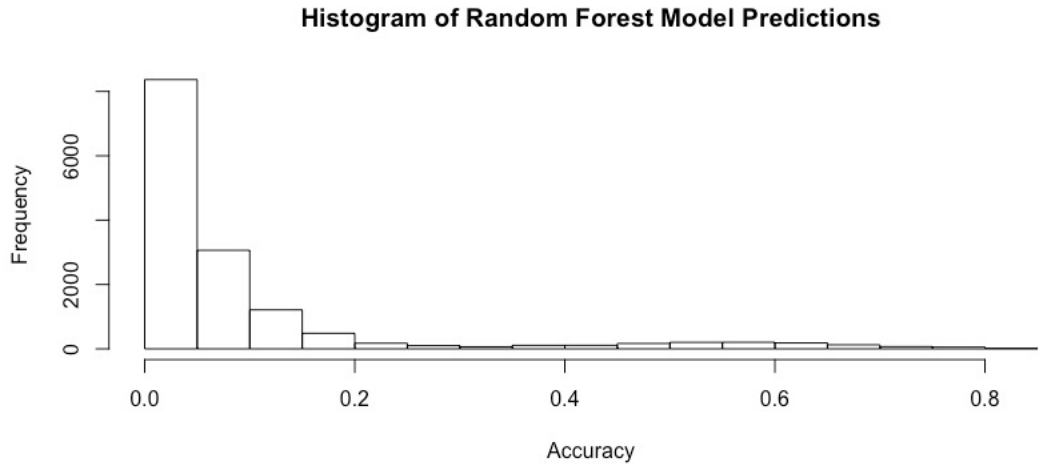

Figure 5: Histogram Plot of Random Forest Model Prediction Accuracies

The most important output of the random forest model is the Variable Importance Plot, presented in Figure 6. Per the Variable Importance plots we see that salepertrans, aka Sales per Transaction, and Pre 2009 Sales are the top 2 most significant variables in predicting RESPONSE16. As we mainly focus on the %IncMSE results, we can see that previous sales and transaction related variables are more important in predicting RESPONSE16 than demographics related variables such as Median Income and Age.

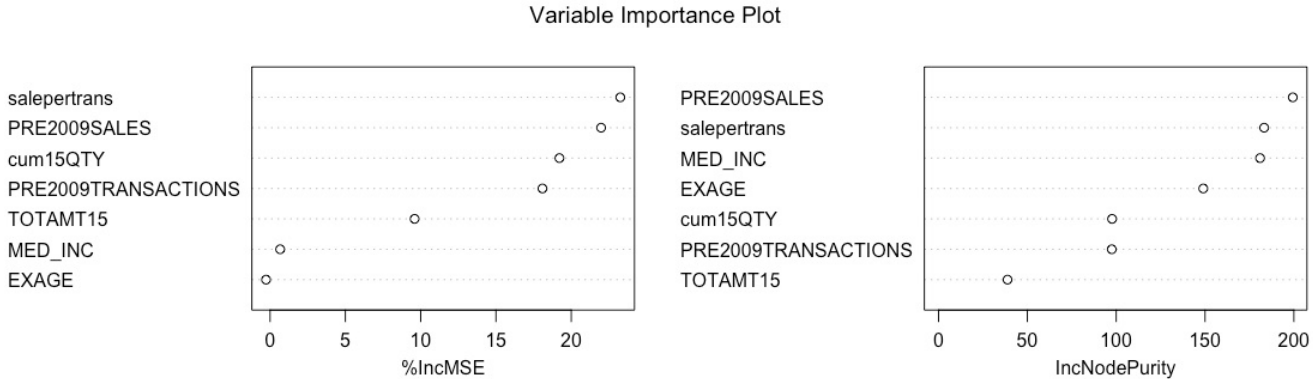

Figure 6: Variable Importance Plot

Per the results obtained through random forest model, which is the better performing model of the two models we conducted, we use the variables salepertrans and PRE2009SALES in order to determine a cut-off value. In order to determine the cut-off point, we compare the mean of the two significant variables mentioned earlier against the mean of predictions with various cut-off points. Per our assessment, review of the histogram plot and variable importance plots, we select salepertrans as the optimum variable to determine the cut-off point based on and after our analysis, we identify the cut-off point as 40%, meaning that we'll be aiming to target 60% of the customers in the subset dataset.

FINANCIAL ANALYSIS & TARGETING

The cut-off point determined in the previous section, 40%, is our starting baseline for the financial model where we compare the current targeting model and our proposed targeting through our new random forest model. For this analysis we use 4 cut-off points in addition to 40%, which are 30%, 35%, 45% and 50%. When these cut-off points applied, we get a higher average revenue per customer than the current targeting model sample size average revenue per customer which is 272. The Figure 7 below presents both the current targeting method and targeting with the new random forest model, as well as comparison between these two targeting approaches in regard to profit.

As we review the results presented in Figure 7, we see that the targeting with the random forest model results in approximately 25% less profit than the current model. The current targeting method yields $3,967,650 in profit whereas with the random forest model we see that the probable profit is at maximum $2,911,080. The $1,056,571

loss in profit is not acceptable and proves that the random forest model failed to offer an optimal targeting different than the current method.

Valuing the Random Forest Models with Alternative Cutoff Rules using the Test Sample (N = 14,695) and $2.00/Mailer Promotion Cost

	Random Forest Model Cutoff Points				
Predicted Probability (percentage)	0.30	0.35	0.40	0.45	0.40
XYZ Current Targeting Methods					
Sample Size (All Customers Get Direct Mailing)	14,695	14,695	14,695	14,695	14,695
Average Revenue per Customer	272	272	272	272	272
Direct mail cost per Customer	2.00	2.00	2.00	2.00	2.00
Ave. Revenue Minus Mail Cost per Customer	270.00	270.00	270.00	270.00	270.00
Profit with of Current Targeting Methods	$3,967,650	$3,967,650	$3,967,650	$3,967,650	$3,967,650
XYZ Targeting with New Model					
Number of Customers Targeted	10,287	9,552	8,817	8,082	8,817
Average Revenue per Customer	285.00	289.00	290.00	298.00	300.00
Direct mail cost per Customer	2.00	2.00	2.00	2.00	2.00
Ave. Revenue Minus Mail Cost per Customer	283.00	287.00	288.00	296.00	298.00
Profit with New Model	$2,911,080	$2,741,352	$2,539,296	$2,392,346	$2,627,466
Profit Increase or Loss with New Model	($1,056,571)	($1,226,298)	($1,428,354)	($1,575,304)	($1,340,184)
Per Customer Profit Contribution or Loss	($71.90)	($83.45)	($97.20)	($107.20)	($91.20)
Number of Customers in Database	30,779	30,779	30,779	30,779	30,779
Estimated Profit Contribution/Lift of Targeting	($2,213,010)	($2,568,508)	($2,991,719)	($3,299,509)	($2,807,045)

Table 3: Financial Analysis and Profit Comparison

Per our analysis we recommend that XYZ company continue to use the current targeting methods, which proves to yield more profit. The decrease in number of targeted customers and increase in average revenue per customer that random forest model outputted do not generate more profit and a better campaign, thus our study shows that the current methods are valid and should be continued.

Aside from our above findings here are some strategic recommendations that we suggest XYZ company further evaluate:

- Limit the mails sent to customers with ‘inactive’ or ‘lapsed’ status and explore the opportunity in targeting ‘active’ customers who haven’t received mail in the 16th mail campaign. This will also help obtain data where more customers have responded the mail campaign which will offer better performing models in the future.
- Continuously record and analyze sales per transaction and sales per customer in order to observe changing patterns.