Capstone Project: Office supply store data analysis

PGD Applied Data Science

Jyoti Sood

Contents

Background

- An office supply store tests a telemarketing campaign on a sample of 16,000 existing business customers.
- The store sells products such as office supplies, desktops, printers, toners, chairs, desks and insurance.
- Using analysis of data sample, the store wants to know whether a customer is likely to make a purchase during the telemarketing campaign and how much a customer is likely to spend.
- Since targeting all customers is expensive, the store expects data insights to predict the the most efficient strategy of targeting customers that would yield maximum profit.

Objectives

- Use past data to profile the customers and identify target market.
- Develop models that will aid in targeting the right customers in future campaigns to yield better profits
- Show the financial value of using models.

Methodology

Methodology

Data exploration

- The original dataset consists 16,173 customers.
- Major chunk of 'null' values in Language (4,500) and Last Transaction Channel (443).
- 3,744 uncategorized employees. The employees categories are relabelled to ordinal data categories small, medium, large, big, huge, unknown.

Feature Engineering

- Date of First Purchased dropped and new column created to denote length of membership.
- Customer number column dropped.
- Instead of throwing off data with null values in 'Language' and 'Employees', it was imputed as 'Unknown'.
- In Historical Sales Volume and Campaign Period Sales - negative values was replaced with zero. Also these features were scaled and transformed as the data was skewed.
- A new binary data column from Campaign Period sales created - 'Purchase' denoted by '1' and 'Non Purchase' by 0.
- Converted categorical variables to dummy variables.

Model Creation and Analysis

Two Models created:

- Model I : Grid Search CV with XG Boost Classifier was created to predict whether a customer makes a purchase during campaign or not.
- Model II: XGBoost Regressor was used to identify the most important features that affect the sales volume ($ Transaction Size) and to predict sales volume and estimate profit

The variables identified by these models were used to estimate profit
E(Profit)=.22*Prob(Sale)*Est(Transaction Size)-$8.40*Prob(Sale)-$45.65

Analysis

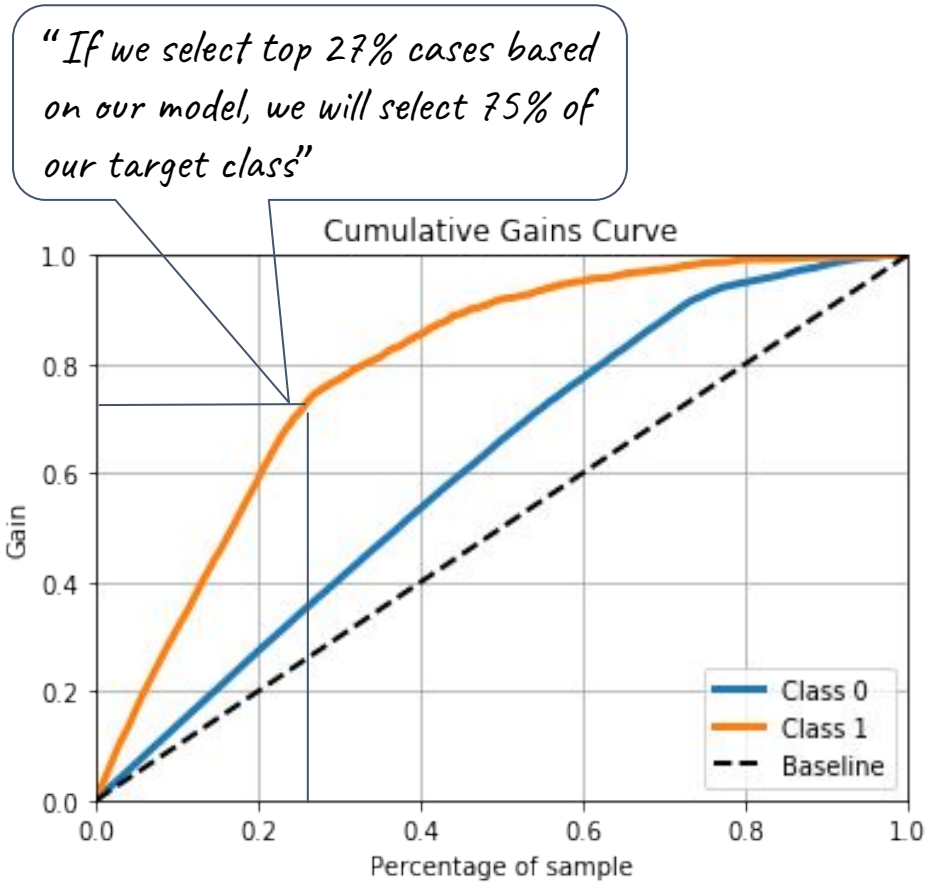
Significant Variables Identified by Ist (Classification) Model :

- Length of Membership
- Historical Sales Volume
- Number of Prior Year Transactions
- Number of Employees: Unknown
- Purchased Monitor
- Purchased Standard Chair
- Number of Employees : Small
- Transaction Channel Used:Mail
- Purchased Office Supplies
- Purchased Printer

•Significant Variables Identified by IInd (Regression) Model:

- Historical Sales Volume
- Length of Membership
- Number of Prior Year Transactions
- Number of Employees : Unknown
- Repurchase Method: Notice
- Purchased Standard Chair
- Number of Employees: Huge
- Transaction Channel Used :Mail
- Purchased Monitor
- Purchased Computer

Cumulative Gains Chart

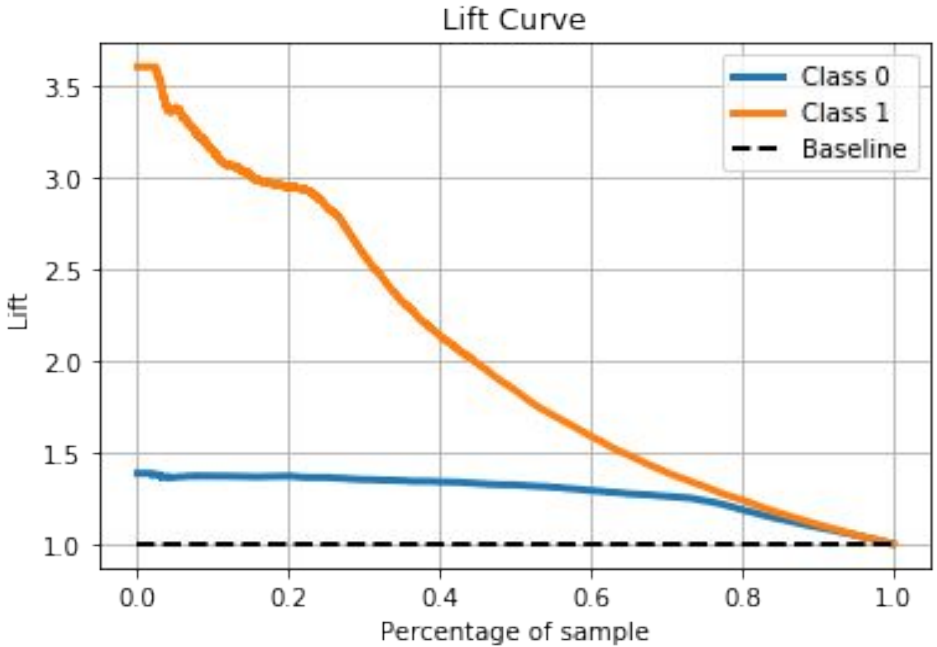


- Cumulative Gains chart depicts cumulative of percentage of Actual buyers (**Cumulative Actuals**) on Y-Axis and Total population on X-Axis in comparison with random prediction (**Gains Chart Baseline**), shown with a dotted line.

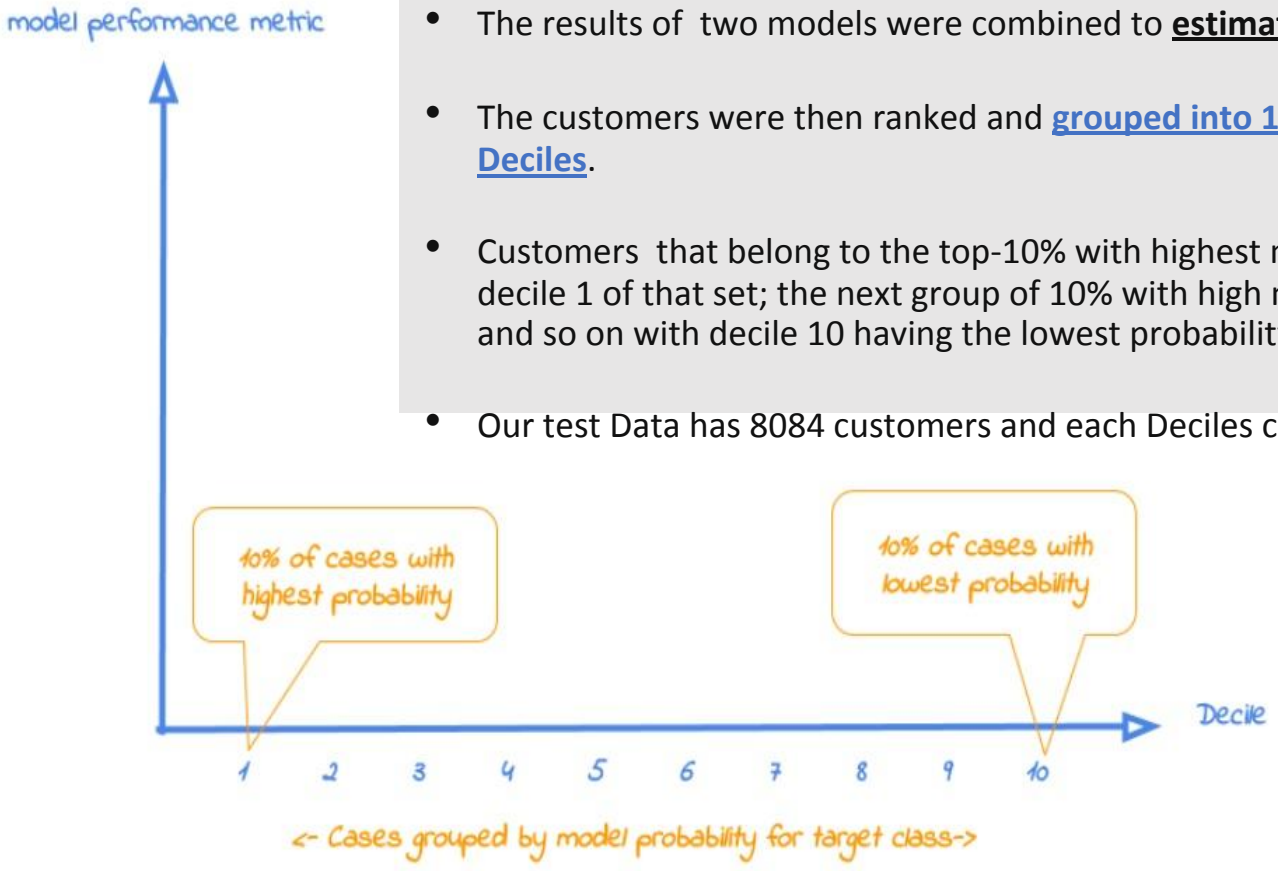
- The cumulative Gains chart here shows, that by the time we covered 27% of the population, we identified 75% of the buyers and by reaching 60% of the population identified 90% of the buyers.

Lift Chart

- Greater the area between lift curve and baseline, better the model.
- A high lift in our model means that during the Marketing Campaign, targeting those customers suggested by our Predictive Model would likely get **3.5 times buying response** as compared to targeting customers at random.



Understanding Deciles



- The results of two models were combined to **estimate profit**.
- The customers were then ranked and grouped into 10 equally large groups called Deciles.
- Customers that belong to the top-10% with highest model probability in a set, are in decile 1 of that set; the next group of 10% with high model probability are decile 2 and so on with decile 10 having the lowest probability.
- Our test Data has 8084 customers and each Deciles consist of average 808 customers.

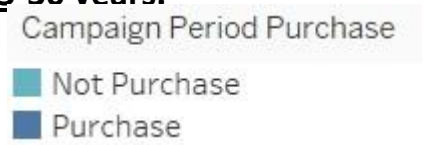
Profit

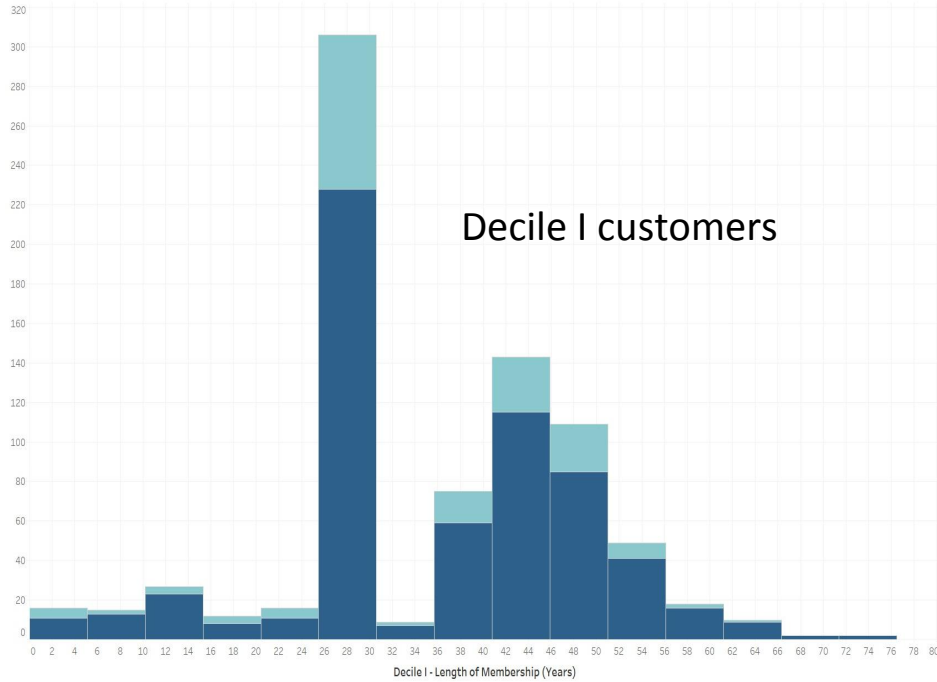
- Overall, the marketing campaign led to **a loss of $14.50 per person**
- Only first Decile was profitable, with **a $167,000.00 profit** and a forecasted profit of **$2.20 M over a customer base of 100,000.**

Tenure of Loyalty with the Store

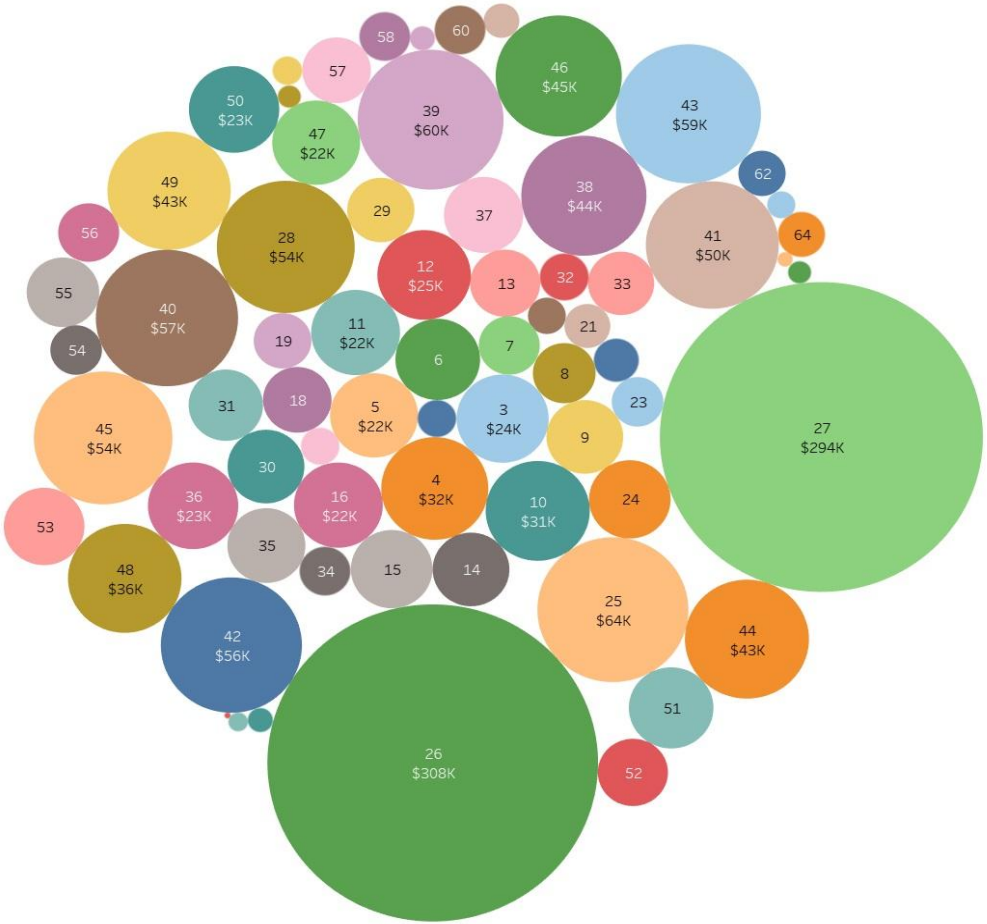
- Most number of customers who buy during campaign, were loyal to the store **since last 26-32 years**.
- **Loyal customers buy more** as compared to new customers.

- In Decile I, customers who purchased the most during campaign were **loyal since 25-30 years.**



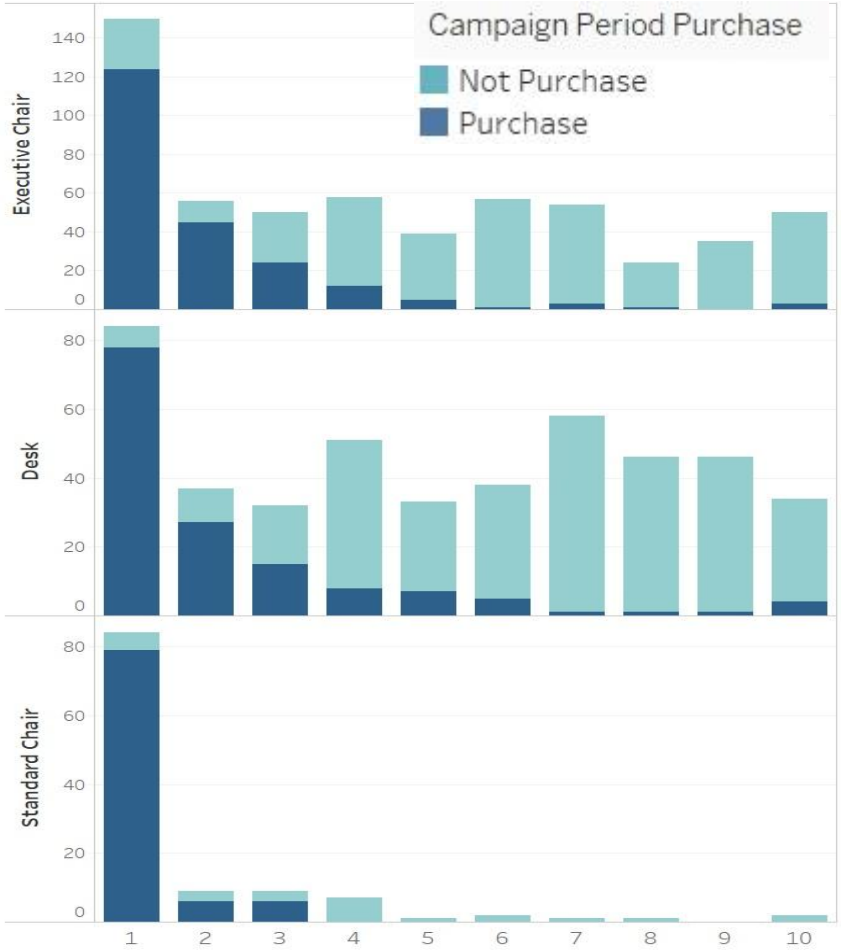


Campaign Period Sales and Length of Membership



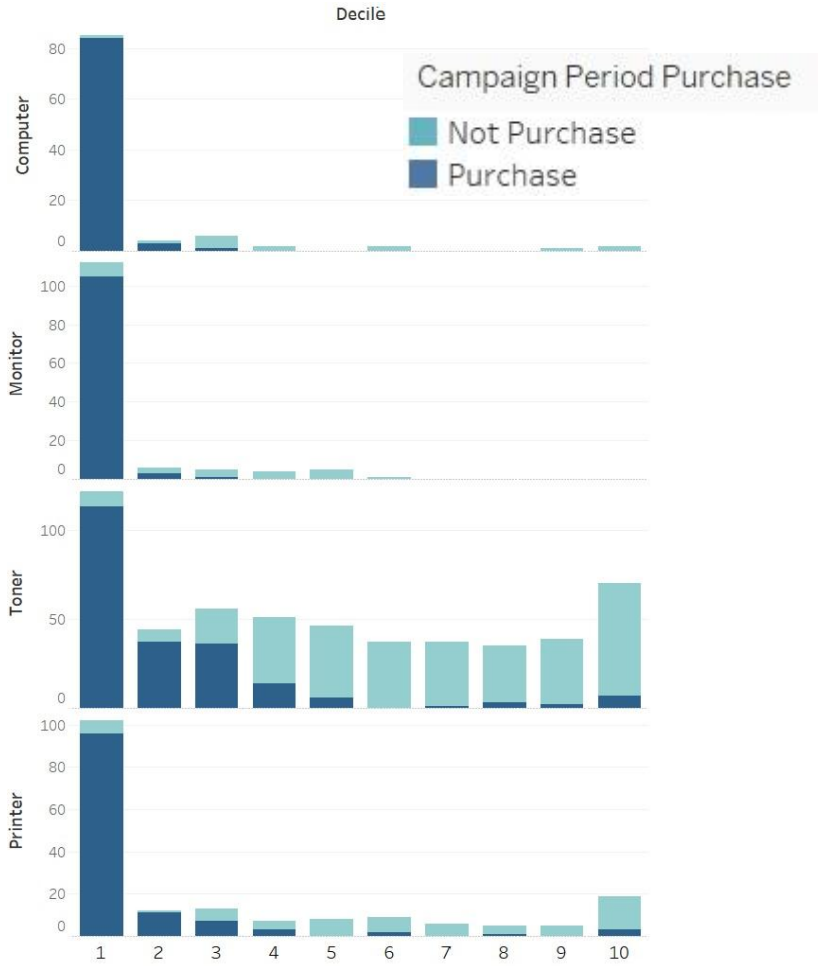
Larger sales volume during the Campaign period were observed in the clients who have **been loyal to the company for 25 years** and above.

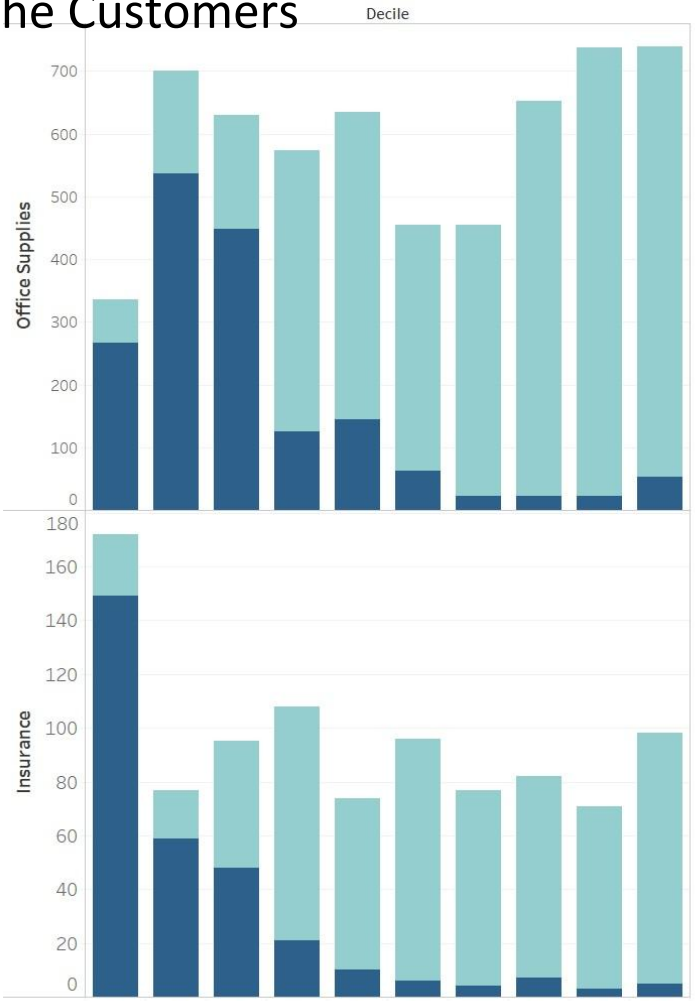
Insights on Products Purchased by the Customers



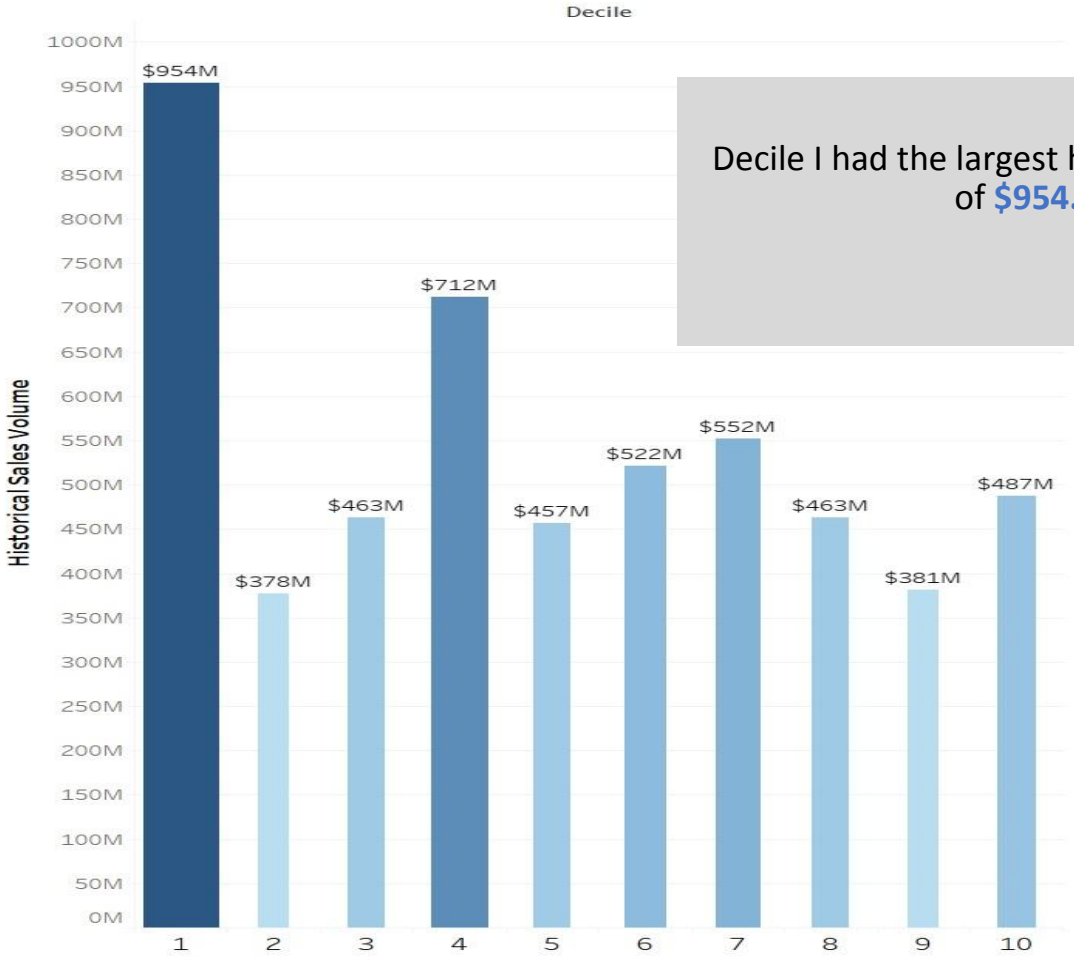
- Decile I customers who made purchase during campaign period had prior history of buying:
  - **85% Computers** of all sold
  - **82% monitors** of all sold
  - **79% Standard Chairs** of all sold
  - 63% Printers of all sold
  - 40% Toners of all sold
  - 40% Executive chairs sold
  - 35% Desks of all sold
  - 32% Insurance of all sold
  - And **only 28% of all Office supplies sold**

Continued - Insights on Products Purchased by the Customers





Historical Sales Volume ($)



Decile I had the largest historical sales volume of **$954.00 M**.

Who are the target customers?

Profile of Decile 1 Customers

- Have been loyal for **35 years (**average).
- Does **16 previous transactions (**average).
- $953 M in historical sales (17% of Total) volume and $1.21 M sales during campaign.
- Are almost exclusive buyers in case of Monitor, Printer, Computer and Standard chair.
- Except Office supplies, these customers are leading buyers of all other products.
- **Mail** is their **most used** transaction channel and **Repurchase** is mostly done after a **Notice**.
- These are not small employee businesses.

Profile of customers not in Decile 1

- Have been **loyal for 26 year (**average).
- **14 previous transactions** (average).
- All other deciles put together make a purchase of mere $0.730M during campaign period.
- **Office supplies** are their main buy.
- These businesses are small, having 1-10 employees.

450% ROI delivered :Using Model in Targeted Marketing

Without Model - Targeting everyone in list

1. Total customers in Test Set= 8,084
2. **Average campaign cost** for per business contacted is $45.65 and total cost for reaching all in the test set is **$369,035.00**
3. Projected **loss** per customer is - $14.40 (negative)
4. Projected **Loss** to be incurred in test set: -$116,491.00
5. Return on Investment (ROI) % : **-31.50**

vs

Targeting only to people in the top profitable Decile of the Model

1. Total customers in Decile 1= 809
2. **Campaign cost** per business is $45.65 and total cost for reaching customers in this Decile **is $36,930.85**
3. **Projected revenue** from per customer is **$ 206**
4. Expected **Profit** to be incurred in First Decile is **$166,448.00**
5. Return on Investment (ROI) %= 450.00

Scenario:

- Client wants to runs a similar marketing campaign to reach 16,000 customers.
- Using the model, client would only target top 27% (4,320 ccustomers) and could make a **profit of $889,920.00**
- However, targeted all the 16,000 customers at random, would **lead to a loss of $230,560.00**

Conclusion

- Marketing to list of customers identified by our models **increases** buying rate by a multiple of **3.5X** (of baseline response rate).
- Initial data analysis revealed that Office Supplies are the most sold product of the store.
- However, the model identified that our core customers (decile 1 customers) who made most purchases in campaign period had **prior history of purchasing computers, monitors, printers and standard chairs (not office supplies).**
- During the campaign, customers who had been loyal for 26/27 years, were the **top buyers**.
- Historical sales volume and Number of prior transactions are significant variables.
- Average number of **prior transactions is 15**.
- All customers are disinclined to be solicited by Email.
- Repurchase method most popular among all customers is 'Notice'.

Recommendations

- It is recommended that **this model be used for future Marketing Campaigns**. Introducing this model could increase ROI to 450 %, from original ROI of -31% (baseline) by targeting selected customers instead of all.
- Number of Employees who are ‘Unknown’ makes a sizeable chunk of first Decile. In the future it will be crucial to collect this data from customers, to be able to target them better. It is recommended **to hire an analyst to update the customer profiles, refine the model and use it on larger customer datasets.**
- Regular promotions, discounts and **cross selling** to customers of Decile 2 and 3 could encourage frequency of buying.
- Non physical transaction channels like Mail, auto-renew and web are the most popular transaction channel across all customers. Client should consider investing in upgrading these channels.
- Loyalty is a significant variable. It is therefore important to target old customers. Tier based ‘Loyalty Programs’ could be introduced. Members could be rewarded through points or ‘$’ Coupon rewards for being a loyal customer.

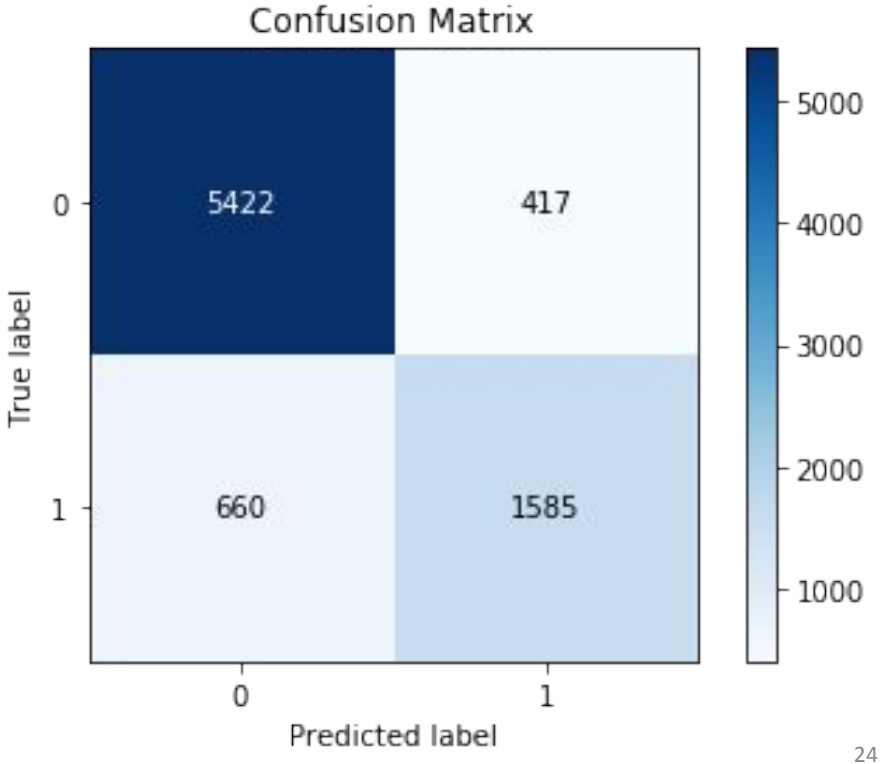
Appendices

Model I- Classification Model

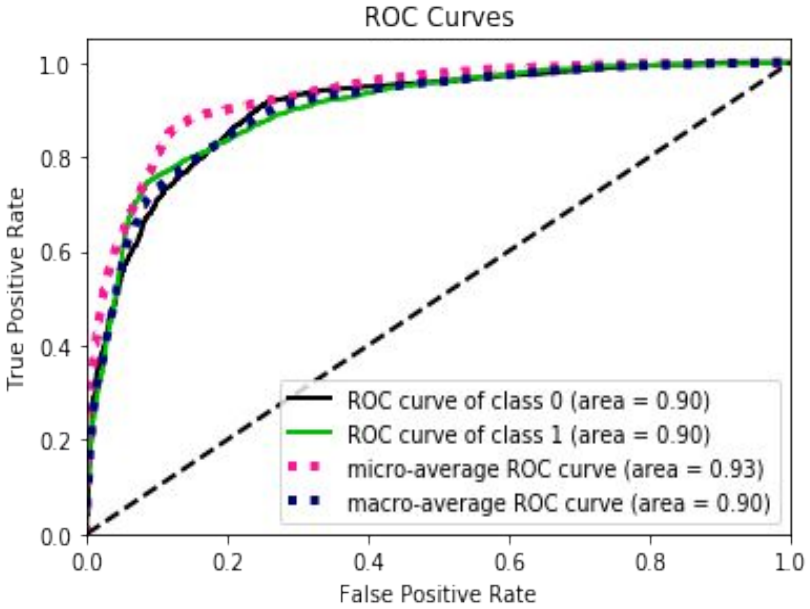
- Since finding whether a customer made a purchase or not during the campaign period is a binary classification problem, the first model was a classification model .
- Training - Model trained on training set and then tested on validation set.
- The target variable was y (whether customer made a purchase, 'yes' or 'no' during the marketing campaign).
- The following machine learning models were trained and tested.
    - Logistic Regression (Accuracy Score 72%).
    - Random Forest Classifier (Accuracy Score 83%).
    - Finally , Grid Search CV with XGBoost Classifier was selected (Accuracy score of 92.7% on Training data and 90.11% on Test Data)
- Lift and cumulative gain curves were calculated.

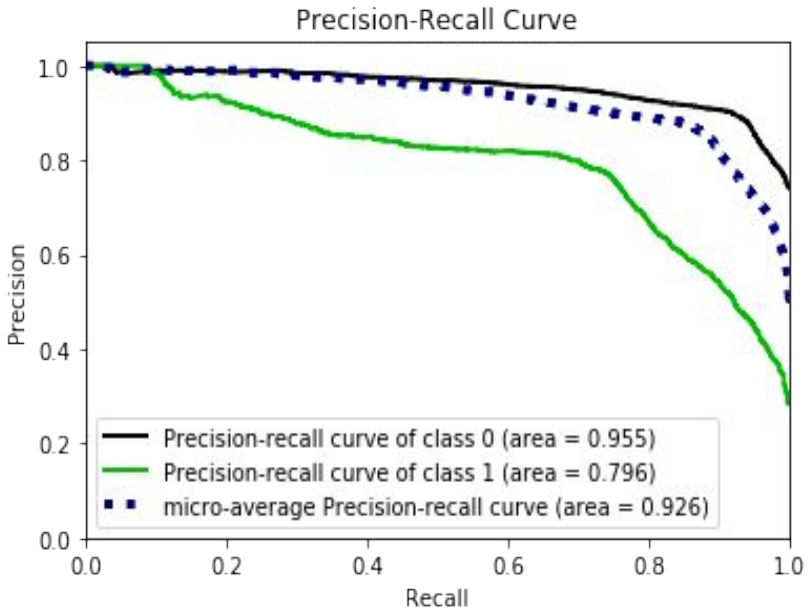
Classification Model's Accuracy

- The confusion matrix of the model indicates that only 660 out of 8084 (8%) values which made purchase were misclassified as non buyers
- Only 5% which had not made a purchase were classified as buyers.
- It indicates that **model is robust** and predicts accurately about customer purchase more than 92% of the times.

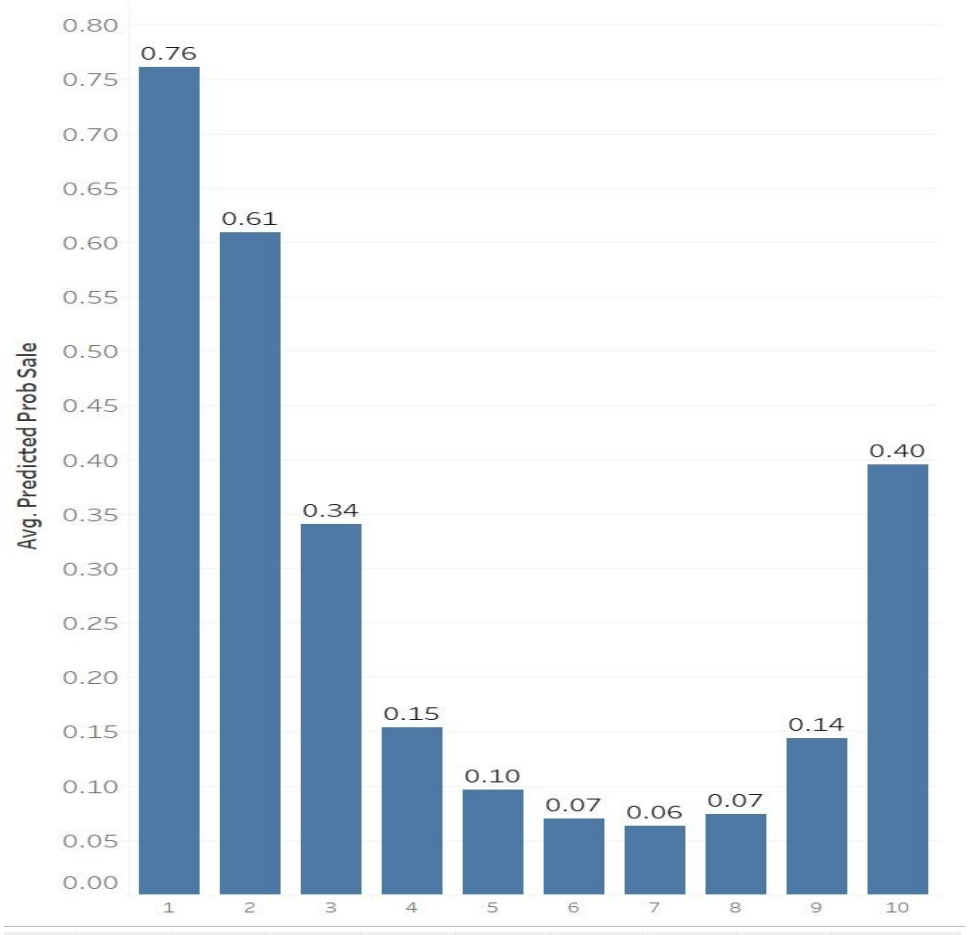


Accuracy Metrics for Classification Model





- The ROC curve is also a useful indicator of how well the model is able to perform classification. The high area under the curve indicates better accuracy of the model.
- High scores for both curves in PRC cuve show that **the classifier model is returning accurate results (high precision), as well as returning a majority of all positive results (high recall**).



Probability Predictions by classification Model whether a customer will make a purchase or not, categorized by each Decile.

Model II: Regression Model

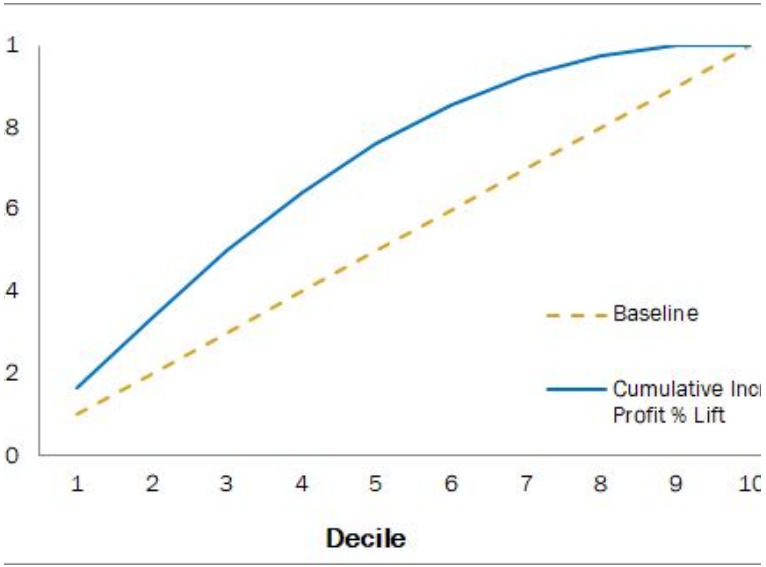
- Regression Model created to complement the classification model. It would **predict the quantity of sales ($)** during the marketing campaign.
- Data split equally in ratio of 50:50 into train and validation sets.
- XGBoost Regressor was used to identify the most important features that affect the sales volume and to eventually help estimate **profit**.
- Model Performance Metrics: R2_score: 0.580 , RMSE for test data 453.005094, RMSE for training data: 310.839843
  .

Gains Chart

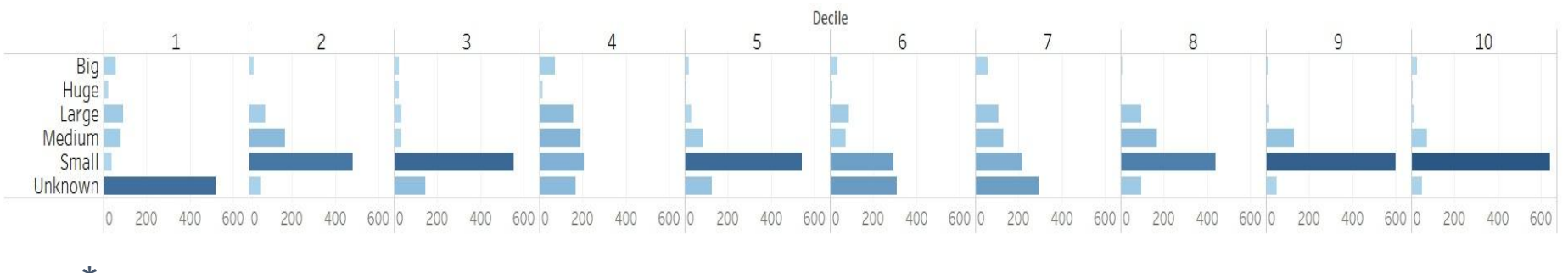
Decile	No. of Customers	Profit per Customer	Lift over Average	Total Profit	Incr. Proj Profit 100k Cust Base ($K)	Total Proj Profit 100k Cust Base ($K)
1	809	$ 206	$ 221	$ 166,898	$ 2,207	$ 2,063
2	808	$ (8)	$ 7	$ (6,210)	$ 67	$ (77)
3	808	$ (28)	$ (13)	$ (22,529)	$ (135)	$ (279)
4	809	$ (42)	$ (27)	$ (33,743)	$ (273)	$ (417)
5	808	$ (44)	$ (30)	$ (35,725)	$ (298)	$ (442)
6	808	$ (45)	$ (31)	$ (36,485)	$ (307)	$ (452)
7	809	$ (46)	$ (31)	$ (36,978)	$ (313)	$ (457)
8	808	$ (46)	$ (31)	$ (37,058)	$ (315)	$ (459)
9	808	$ (46)	$ (32)	$ (37,138)	$ (315)	$ (460)
10	809	$ (46)	$ (32)	$ (37,417)	$ (318)	$ (463)
Total	8,084	$ (14.4)		$ (116,385)	$ -	

Regression Model's Cumulative Gains Chart

Decile I has a $221 lift over average profit. The cumulative gains chart shows that Decile I achieves 17% of the profit, by Decile 3, 50% of profit is achieved and by Decile 6, 85% profit is achieved.



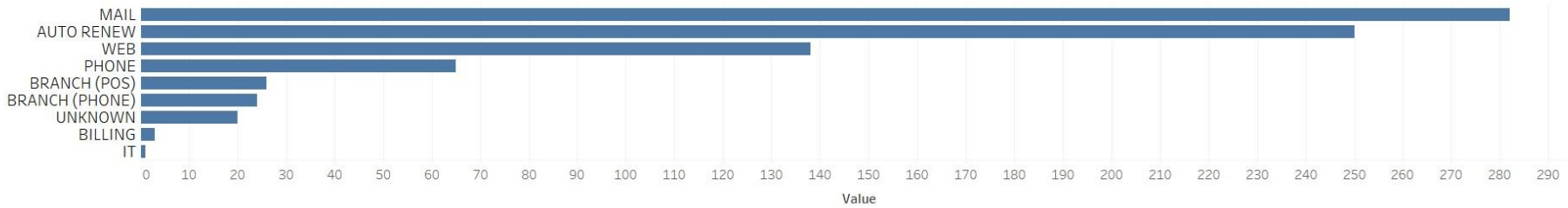
Number of Employees in each Decile



**1-5: Small, 6-10: Small, 11-50 Medium , 51-100: Large, 101-500:Big , 500+: Huge, Missing Data : Unknown

- Number of Employees which were imputed to the label ‘Unknown’ due to missing data, are the largest percentage of customers in Decile I.
- In other Deciles, customers who buy the most are small businesses with 1-10 number of employees .
- Companies in following categories made the most purchases during campaign period:
  Small (27%)
  Unknown(27%)
  Huge (25%)

Transaction Channels used by customers in Decile I



- In Decile I and across all deciles, non branch channels of transaction such as Mail, Auto Renew and Web are more popular than branch (physical) channels. IT is least preferred channel.

Repurchase Methods of customers in Decile I

- **Notice** is the **best Repurchase Method** in first Decile, followed closely by Auto Renew. All deciles follow the same trend.