OFFICE CORP

OFFICE SUPPLY STORE DATA ANALYSIS

UMUT CEMAL YETGIN

Agenda

- Background
- Objectives
- Approach
- Analysis (customer profile)
- Analysis (modeling)
- Recommendations
- Appendix

Background

Campaign

- Initiated telemarketing campaign with 16k customers
- Had sales with around 4k customers during campaign
- Recorded details about each customer that contacted during the campaign

Financials

- Gross margin on sales: 22%
- Campaign cost: $45.65 per business contacted
- Transaction cost: $8.40 per transaction

Objectives

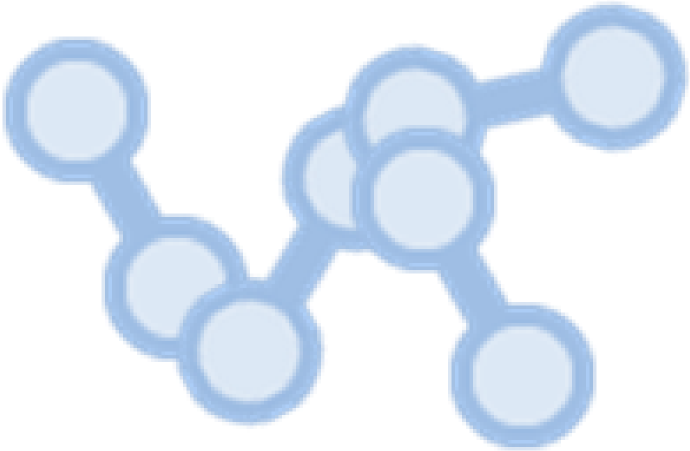
Profile the **customers** that were contacted during the campaign

Develop models to target customers efficiently in future campaigns

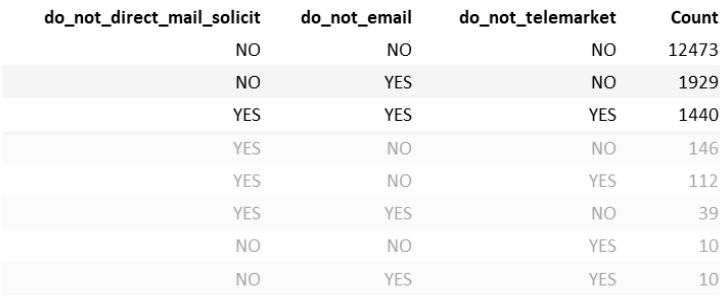
Provide **expected value** of the models with financial benefits

Approach

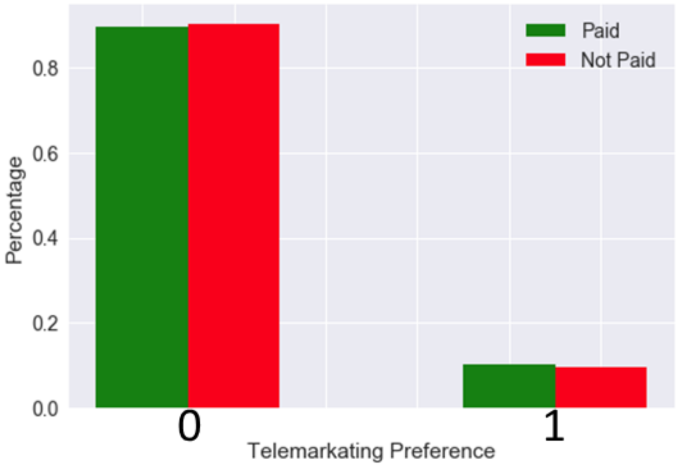
- Retrieved data are cleaned up with respect to the assumptions mentioned in Appendix A
- Customers are profiled with exploratory data analysis (EDA)
- Most important features are picked as predictor variables in models
- Best performing models are selected for classification (purchase probability) and regression (estimated transaction size)
- Profitability is estimated for each customer for future campaigns and illustrated with a lift chart
- Customers are divided into percentiles with respect to estimated benefit and most profitable percentiles are addressed
- Target customers and recommendations on new campaign are shared as a conclusion



EDA – Communication Preferences



do_not_direct_mail_solicit	do_not_email	do_not_telemarket	Count
NO	NO	NO	12473
NO	YES	NO	1929
YES	YES	YES	1440
YES	NO	NO	146
YES	NO	YES	112
YES	YES	NO	39
NO	NO	YES	10
NO	YES	YES	10

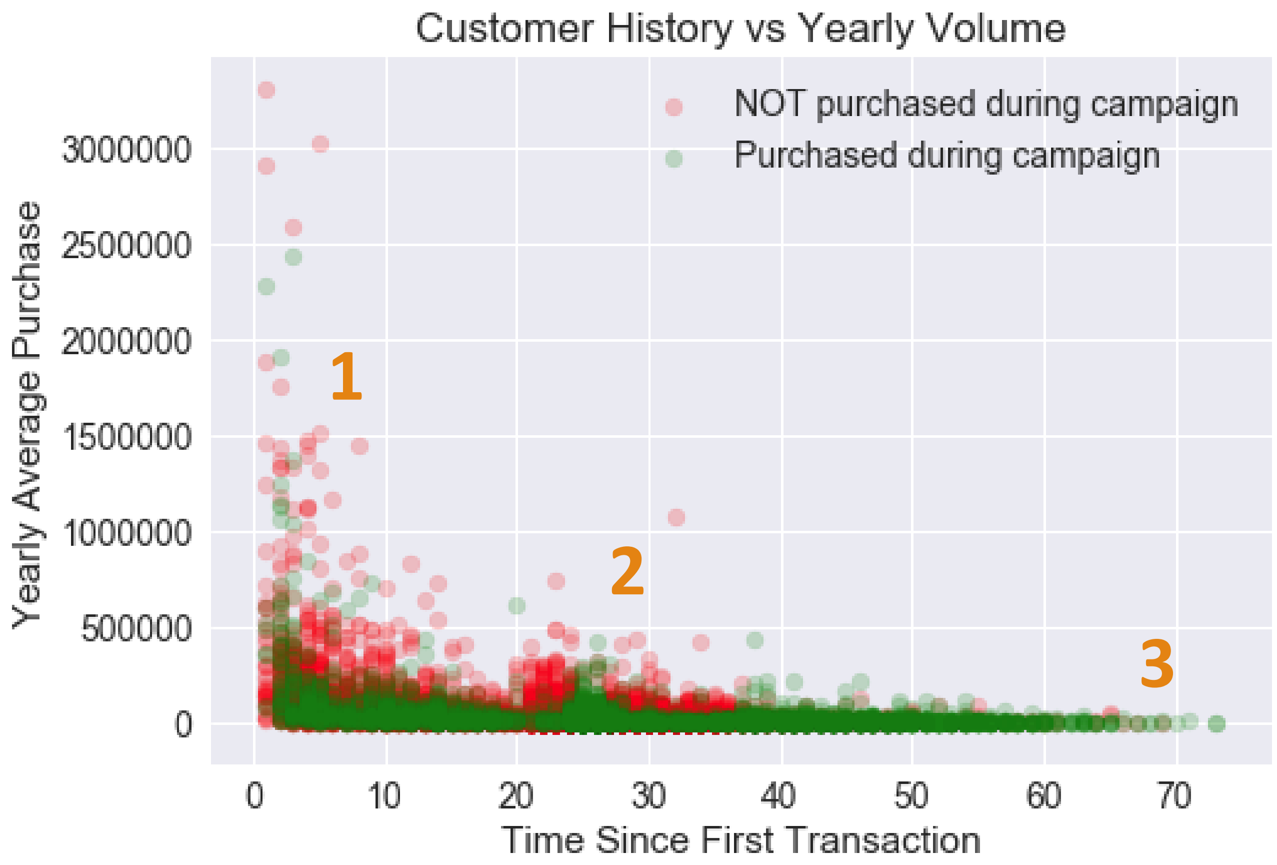


- Most of the customers prefer to be communicated through any channel
- If a customer prefers not to be communicated, rejects all communication channels usually
- For specific channel selection, email is the least preferred
- Customers who purchased during the campaign has similar telemarketing preference with the ones that did not purchase during the campaign

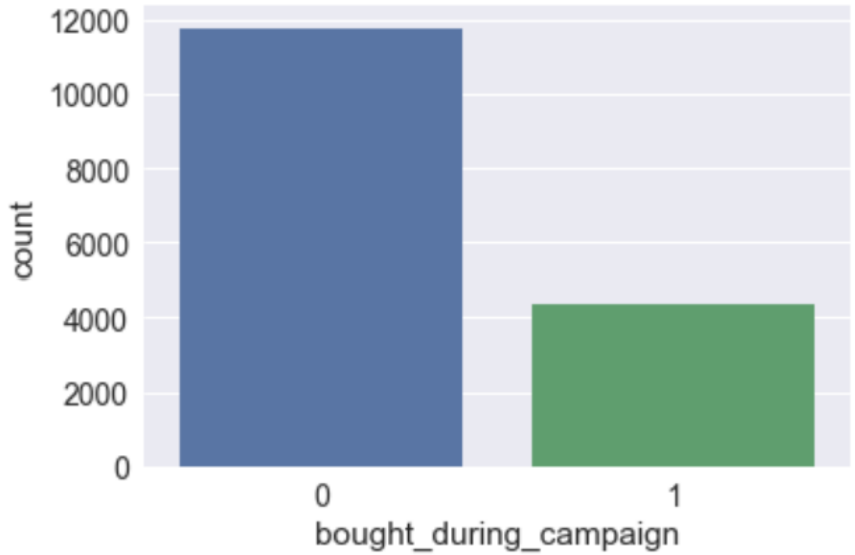
EDA – Historical Patterns

- Yearly average purchase vs customer age can indicate customer attitude

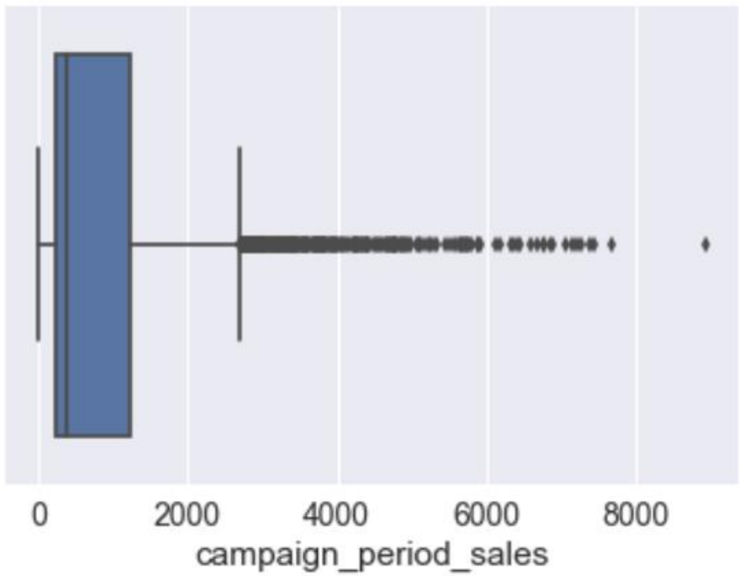
1. New Customers with Intention to Expand Rapidly
2. Loyal and Profitable Customers
3. Old Customers with Less Profit

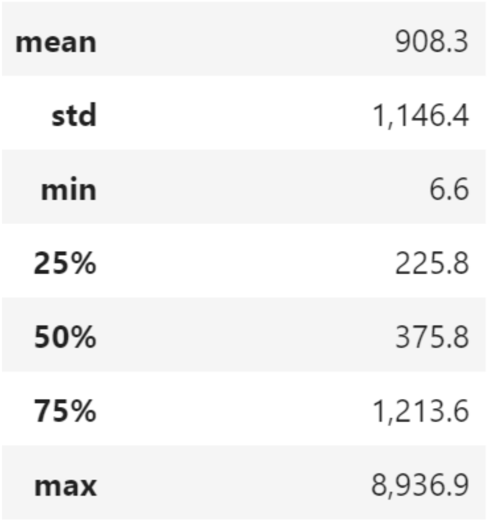


EDA – Campaign Period Sales



Sales Volume for Paying Customers





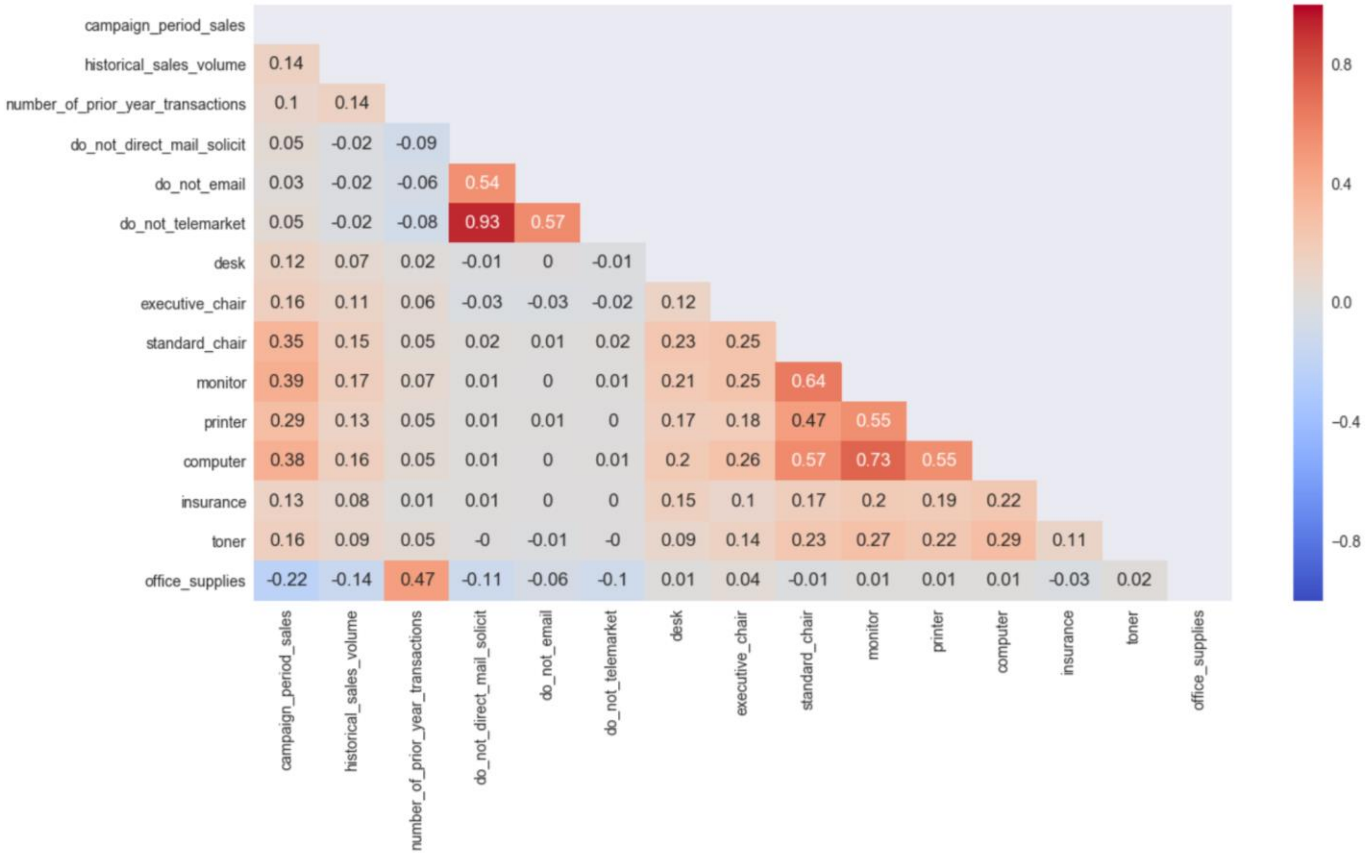
mean	908.3
std	1,146.4
min	6.6
25%	225.8
50%	375.8
75%	1,213.6
max	8,936.9

27% of the customer purchased during the campaign

Some significant volumes, yet sales mostly below $1,200

EDA – Correlations

- Customers communication preferences mostly match for each channel
- Computer, printer, monitor and standard chair are purchased together
- These four items have a positive correlation with campaign period sales, too
- Number of prior year transactions are positively correlated with office supplies purchase

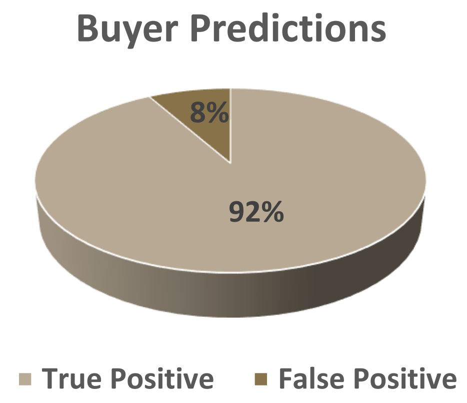


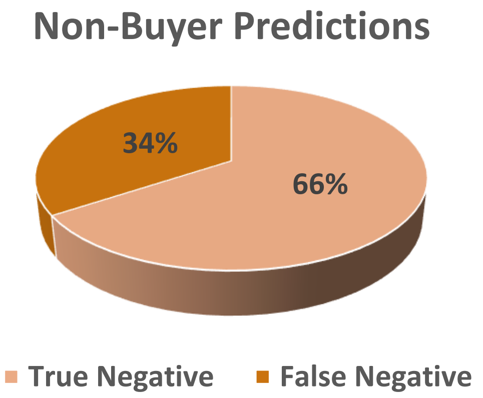
Classification – Predicting Purchase Probability

Independent Variables Selected as Input to the Model

- Years Past since First Transaction (FI = 0.5)
- Yearly Average Purchase (FI = 0.2)
- Historical Sales Volume(FI = 0.1)
- Number Of Prior Year Transactions (FI = 0.1)
- Employee Count (FI = 0.1)

FI = Feature Importance



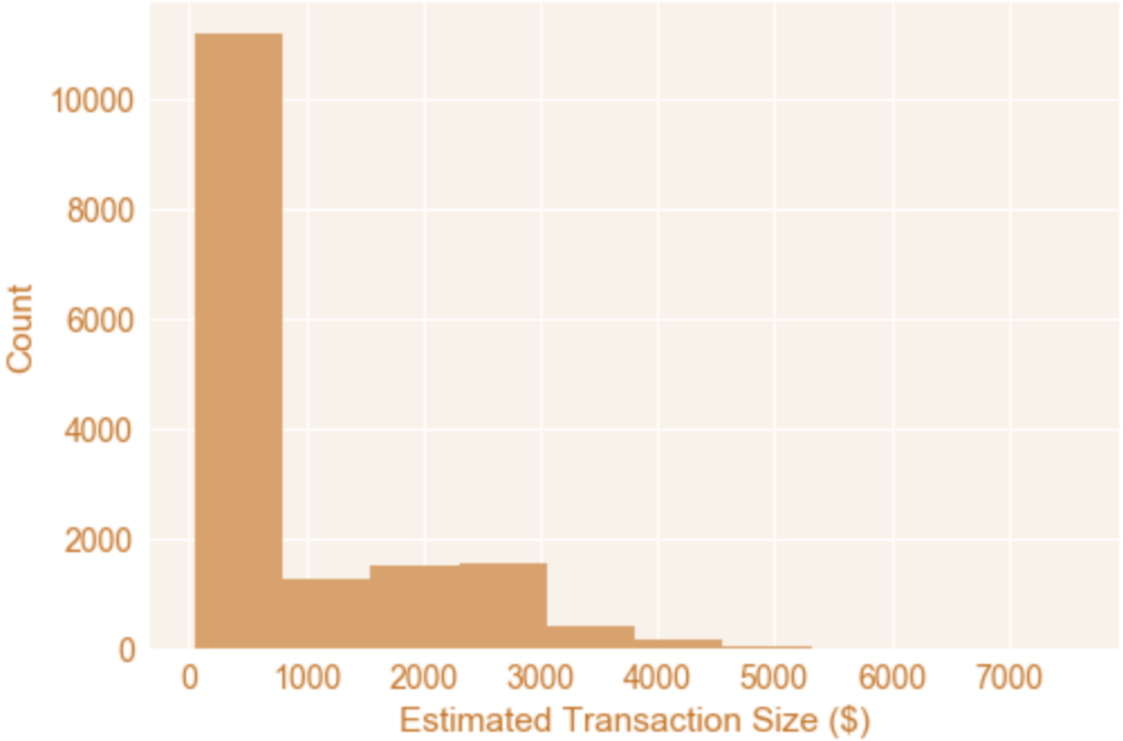


More details under Appendix B

Regression – Predicting Transaction Size

Independent Variables Selected as Input to the Model

- Historical Sales Volume
- Number Of Prior Year Transactions
- Standard Chair Purchase
- Monitor Purchase
- Office Supplies Purchase
- Years Since First Transaction
- Repurchase Method
- Last Transaction Channel
- Employee Count
- Language Unknown or Not
- Yearly Average Purchase

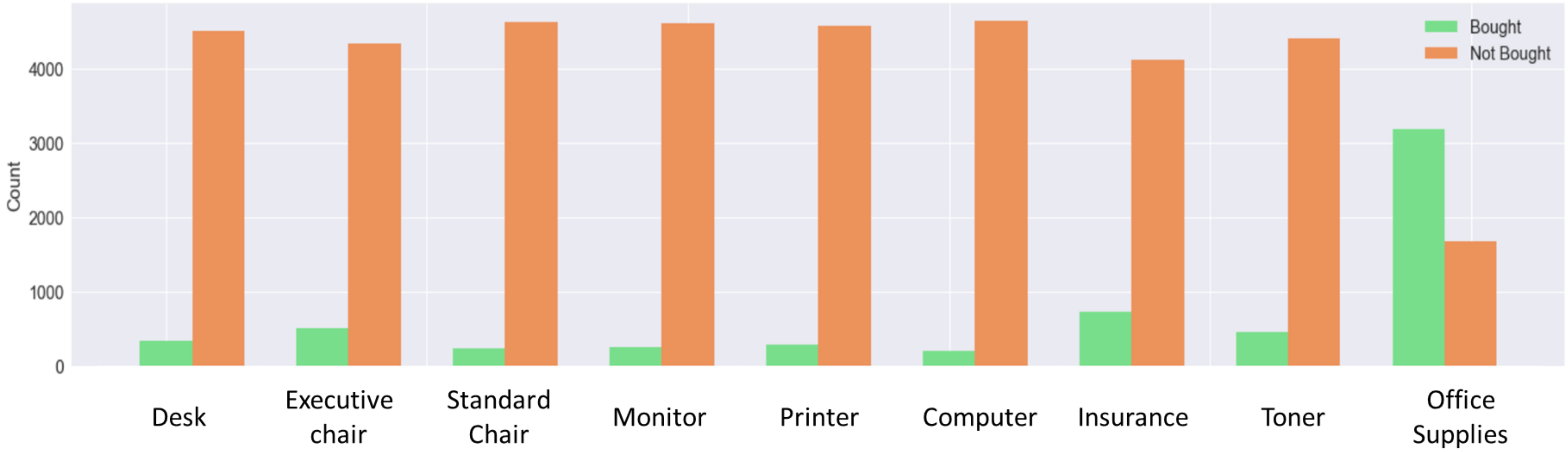


More details under Appendix B

Lift Chart

Decile	Number of Customers	Actual Profitabilty per Customer	Lift over Average	Total Profit	% of Profit	Incr Proj Profit 100k Cust Base($K)	Total Proj Profit 100k Cust Base($K)	Cuml Incr Proj Profit 100k Cust Base($K)	Cuml Total Proj Profit 100k Cust Base($K)
1	1616	$ 295	$ 288	$ 477,090	412%	$ 2,881	$ 2,952	$ 2,881	$ 2,952
2	1616	$ 47	$ 40	$ 75,783	65%	$ 397	$ 469	$ 3,278	$ 3,421
3	1616	$ 6	$ (1)	$ 10,429	9%	$ (7)	$ 65	$ 3,271	$ 3,486
4	1615	$ (16)	$ (23)	$ (25,784)	-22%	$ (231)	$ (160)	$ 3,039	$ 3,326
5	1616	$ (34)	$ (41)	$ (54,175)	-47%	$ (407)	$ (335)	$ 2,632	$ 2,991
6	1616	$ (45)	$ (52)	$ (72,370)	-62%	$ (520)	$ (448)	$ 2,113	$ 2,543
7	1615	$ (46)	$ (53)	$ (73,725)	-64%	$ (528)	$ (457)	$ 1,585	$ 2,087
8	1616	$ (46)	$ (53)	$ (73,770)	-64%	$ (528)	$ (457)	$ 1,056	$ 1,630
9	1616	$ (46)	$ (53)	$ (73,770)	-64%	$ (528)	$ (457)	$ 528	$ 1,174
10	1616	$ (46)	$ (53)	$ (73,770)	-64%	$ (528)	$ (457)	$ -	$ 717
Total	16,158	$ 7.2	$ -	$ 115,937	100%	$ -		$ -	

Decile Analysis – Purchased Items



Office supplies has significantly higher purchase rate; whereas, insurance and toner have slightly higher purchase rate compared to other items

Recommendations

- Company should target customers in first three percentiles for maximum profitability
- Company should proactively offer a pack of computer, printer, monitor and standard chair in case that a customer orders at least one of these items
- Next campaign can focus on office supplies primarily and insurance secondarily
- The company should consider continuous engagement of a data science team to further enhance profitability of such campaigns

Thank You

Appendix A - Preprocessing

- Payment plan row is dropped from repurchase method since there is only a single entry
- Below column nan values are included in analysis as ‘Unknown’
  - last_transaction_channel, number_of_employees, language
- Below column nan values are dropped from the analysis
  - campaign_period_sales, historical_sales_volume, date_of_first_purchase, number_of_prior_year_transactions, do_not_direct_mail_solicit, do_not_email, do_not_telemarket, repurchase_method, desk, executive_chair, standard_chair, monitor, printer, computer, insurance, toner, office_supplies
- YY values are updated as Y with the assumption that they stand for the same meaning
- Excluded negative campaign sales (product returns) and negative historical sales volume
- Excluded wrong date value (later than today)

Appendix B – ML Activities

Classification - Key Actions

- Data split with **80/20** ratio for training and testing data
- Model performance evaluated with **accuracy score**
- Used **Random Forest** algorithm with **Grid Search** for the best performance

Regression - Key Actions

- Data split with **80/20** ratio for training and testing data
- Model performance evaluated with **mean squared error**
- Used **Random Forest Regressor** algorithm with **Recursive Feature Elimination**

Appendix C – Python Code