Officework Supplies

Your one-stop center

Go-Get project

By: Daniel Chow

1 What is Go-Get project?

Background

- Officework Supplies tested on one **telemarketing campaign** recently.
- Test **results** are analyzed to review any insights that can be leveraged for future similar campaign(s).

Our objectives are to :-

- **understand the characteristics of customers** who responded to the campaign;
- leverage the results and **enhance customer targeting** using a 'two-stage' model (i.e. *predict the likelihood to response and estimate sales values*);
- **evaluate and maximise profit** with model vs. random targeting.

1 High Level Approach

Methodology

Get → Cleanse → Explore → Model → Result

Get / Cleanse / Explore

- **Acquire** data
- **Identify** & **remove** any invalid / duplicate records
- **Treat** missing values using standard **data imputation** method
- **Identify** strength of relationship between variables (aka features)
- **Manage** highly correlated variables using **dimension reduction** technique
- **Drop** outliers / anomalies records
- **Create** new feature(s)
- **Analyse** each data variable **and detect** any patterns/trends

Explore / Model / Result

- **Select** & **Transform** features for model development
- **Split** 50:50 data for model development / test
- **Select** appropriate algorithm model(s)
- **Develop** ‘two-stage’ models:-
  - ‘*first*‘ propensity model to predict customer’s likelihood to response
  - ‘*second*’ model to estimate customer’s transaction spend size($)
- **Compute** expected profit from models
- **Rank** profit from highest to lowest values and **categorize** them into 10 equalled sub-groups

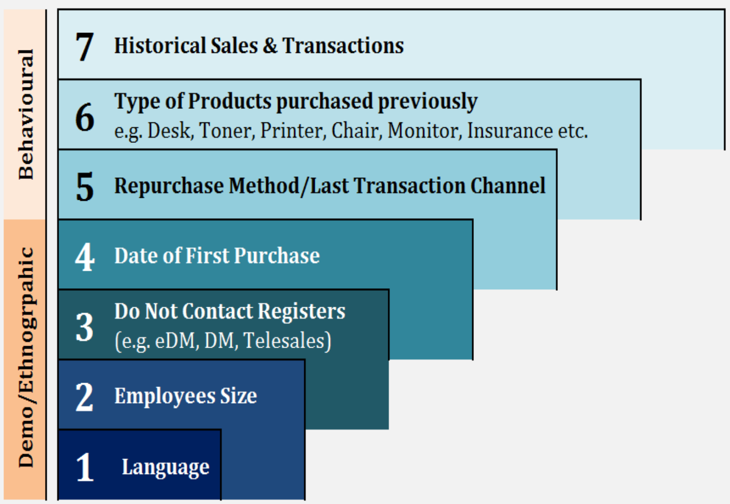
2 Get, Cleanse and Explore

A Quick Glance

What do we know about the data?

1. Test results consist of **16,172** customers.

2. There are **21** features but can be grouped into **7** broad category as below:-



3. **10** customers with negative campaign sales value / historical sales which probably due to refund from previous sale.

4. Approximately **1.5K** customers explicitly do not want to be contacted via Telemarketing but were targeted. Surprisingly, **452** actually responded.

5. Some other findings include :
   i. invalid record / typo errors
   ii. missing values

Features	# of Missing Records	% of Missing Records
Language	4,467	27.64%
Number of Employees	3,744	23.17%
Last Transaction Channel	442	2.73%
Toner	2	0.01%
Insurance	2	0.01%
Printer	1	0.01%
Monitor	1	0.01%
Standard Chair	1	0.01%
Executive Chair	1	0.01%

Next step, data cleaning process!

Note: The term features, variables, dimensions and columns are used interchangeably herein refer to the same object.

2 Get, Cleanse and Explore

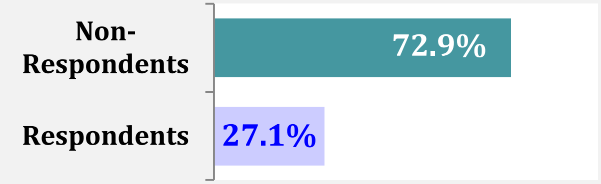
	Step 1	Step 2	Step 3
			
3-STEPS	**DATA INTEGRITY**	**IMPROVE QUALITY**	**INCREASE CLARITY**
DETAILS	1. Fix data errors 2. Remove records with negative sales values 3. Impute missing data as shown in page 5	1. Assess strength of relationship between variables 2. Group/Combine correlated features using dimension reduction process 3. Remove outliers / anomalies	1. Design new features: - Tenure - Product Mix - Contact Channel - Average Sales/Tran 2. Create standard buckets for analysis review: - Historical Sales - Prior Sales Trans - Tenure

2 Get, Cleanse and Explore

**** PERSONAS ****

RESPONDENTS' PROFILES

- lower **Historical Sales**
- higher **Prior Year Transactions**
- longer **Tenure**
- specific **Number of employees size**
- specific **Language**, other than English



For details of dashboard, please refer Appendix

HISTORICAL SALES

Relatively lower historical sales

<= $150K

Historical Sales Band	Relative Index
>= $1.5M	76
$800k- $1.5M	88
$500K- $800K	90
$300K- $500K	99
$150K- $300K	113
$ 50K - $150K	119
< $50K	111

PRIOR YEAR TRANSACTIONS

Relatively higher number of prior year transactions

>= 15

Prior Yr Trans Band	Relative Index
>=40	451
30-39	563
20-29	120
15-19	117
5-14	82
<5	53

TENURE *(length of relationship)*

Relatively longer tenure

>=25 years

Tenure Band	Relative Index
>=50 Yrs	1539
40-49 Yrs	985
30-39 Yrs	47
25-29 Yrs	172
20-24 Yrs	8
10-19 Yrs	79
5-9 Yrs	73
<5 Yrs	101

EMPLOYEES SIZE

Either customers with smaller staff size or very large base

Num of Employees	Relative Index
1-5	112
6-10	93
11-50	114
51-100	61
101-500	115
500+	133

LANGUAGE

Language	Relative Index
Arabic	103
Chinese	54
English	102
French	44
German	95
Greek	123
Hebrew	38
Hindi	36
Italian	54
Japanese	134
Korean	0
Pashto	90
Polish	54
Portuguese	87
Russian	67
Spanish	72
Thai	269
Vietnamese	77

***Relative Index** measures the strength/weakness within each features and compare them between respondents and non-respondents.*

2 Get, Cleanse and Explore

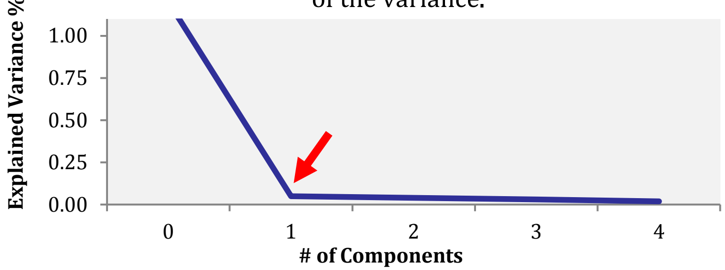
- Some features were **correlated**[1] to each other. Two aspects identified as below.
- These features will distort some of the models (e.g. logistic/linear regression) if untreated.
- Apply dimensions reduction technique to overcome this issue.

Aspect #1 | Products purchased previously

Monitor, Computer, Printer, Standard Chair & Product Mix

Features	Monitor	Printer	Computer	Std_Chair	Prod_Mix
Monitor	**1**				
Printer	0.549	**1**			
Computer	0.727	0.55	**1**		
Standard_Chair	0.640	0.466	0.570	**1**	
Product_Mix	0.625	0.556	0.625	0.573	**1**

Based on the scree plot[2] below, all features can be combined under **1 component** as it explained most of the variance.



All features can be combined into 1 component and each feature contributed approximate equal weight.

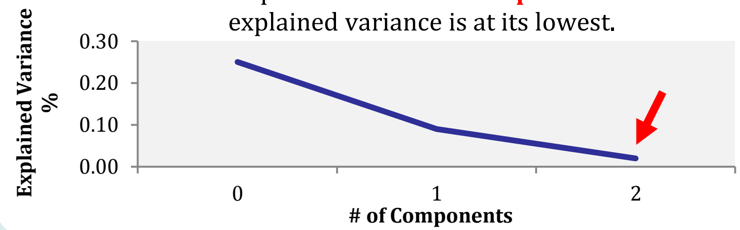
Weights	Monitor	Printer	Computer	Std_Chair	Prod_Mix
PCA_Prt_Mon_Com_Chr	**21%**	**18%**	**21%**	**19%**	**20%**

Aspect #2 | Do Not Contact Register(s)

Do Not Email, Do Not Mail Solicit and Do Not Telemarket

Features	Do_Not_Direct Mail_Solicit	Do_Not Email	Do_Not Telemarket
Do_Not_Direct Mail_Solicit	**1**		
Do_Not Email	0.544	**1**	
Do_Not Telemarket	0.933	0.572	**1**

It is optimal to retain **2 components** as explained variance is at its lowest.



i. **Do Not Mail Solicit & Do Not Telemarket have more weights on first component.**

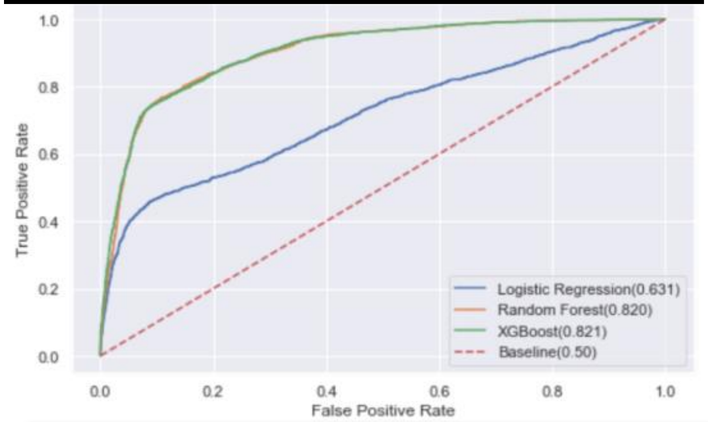
ii. **Do Not Email takes 55% weights on the second component.**

Weights	Do_Not_Direct Mail_Solicit	Do_Not Email	Do_Not Telemarket
PCA_1_DM_Tele	**36%**	**28%**	**36%**
PCA_2_Email	**24%**	**55%**	**21%**

Note1: Correlation measures how strongly one variable depends on another. Value >= 0.55 are correlated.
Note2: Scree plot is used to determine the number of factors to retain in an exploratory factor analysis (FA).

3 'Two-Stage' Model Review – First Propensity Model

Receiver Operating Characteristic [ROC]



SUMMARY Classification Model	Test Dataset	
	Accuracy Score	AUC Score
Logistic Regression	0.785	0.631
Random Forest	0.868	0.820
XGBoost	**0.870**	**0.821**

Ensemble |XGBoost

Accuracy Score on Train : 0.882
AUC Score on Train : 0.838
Log Loss on Train : 4.090

Accuracy Measures on Test

Confusion Matrix :

[[5403 437
598 1509]]

Accuracy Score on Test : 0.870
AUC Score on Test : 0.821
Log Loss on Test : 4.498

Classification Report :

	Precision	Recall	F1-Score	Support
0.0	0.90	0.93	0.91	5840
1.0	0.78	0.72	0.74	2107
accuracy			0.87	7947
macro avg	0.84	0.82	0.83	7947
weighted avg	0.87	0.87	0.87	7947

- Comparison made based on two measures:
  - **Accuracy** - *how well the model predicts the outcome correctly as a total.*
  - **AUC** (aka AUROC) - *how well the model distinguishes positive and negative outcomes , which can be easily visualised in the ROC diagram above. The bigger the area under curve, the better the model.*

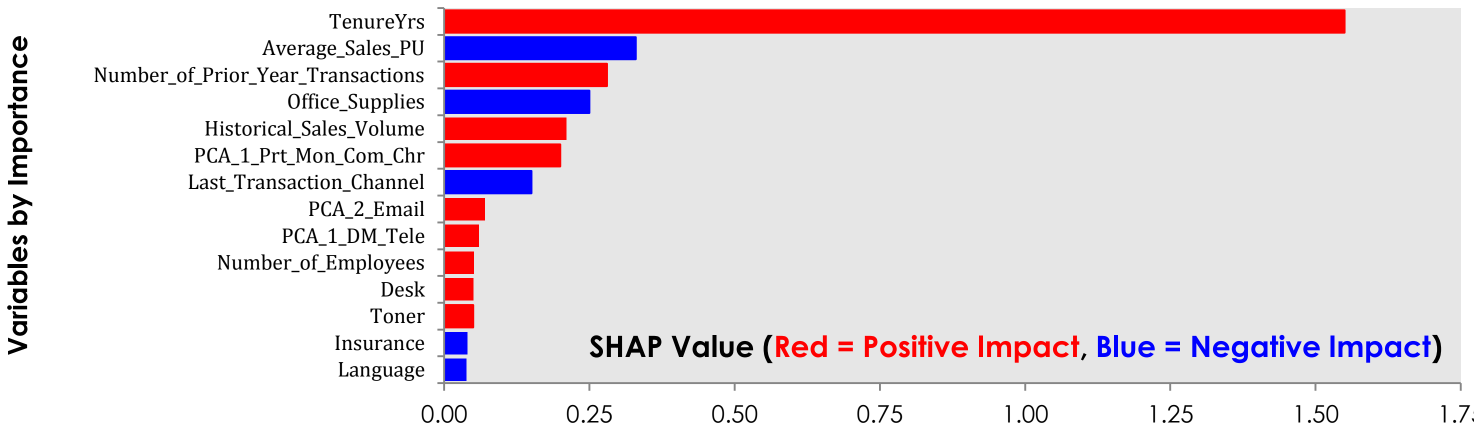
WHY IS THIS IMPORTANT?

- A good model leads to better targeting which can generate more revenue and lowering marketing expenses.
- In this specific case, will focus on AUC score as it helps to maximise conversion rate.

*XGBoost , aka **eXtreme Gradient Boost**, is a decision-tree-based ensemble Machine Learning algorithm that uses a gradient boosting framework.*

3 ‘Two-Stage’ Model Review – First Propensity Model

Result from Best Propensity Model | Variables Importance Plot

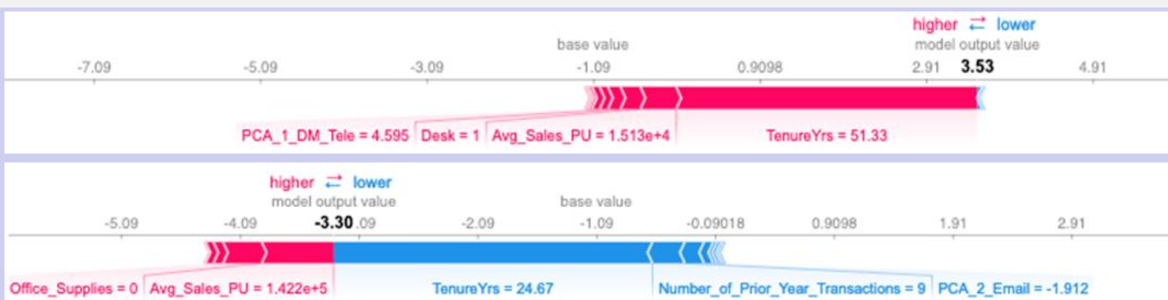


The model takes in each variable based on its significance and calculate the expected value. For example:

- ***Tenure*** *– the longer the tenure, the higher the expected value;*
- ***Number of Prior Year Transactions*** *– the bigger prior year transactions, the higher the expected value;*
- ***Historical Sales Volume*** *– the lower the historical sales, the higher the value;*
- ***Office Supplies*** *– prior purchased of office supplies lower the value.*

Illustration:

Two examples showed how do each variable is calculated towards expected value.



Note: SHAP is known as ***SHapley Additive exPlanations****.*
It helps to quantifies the direction of each significant variables.

3 ‘Two-Stage’ Model Review – Second Regression Model

SUMMARY Regression Model	Test Dataset	
	R2 Score	RMSE
Linear Regression	0.704	0.564
Random Forest	0.803	0.460
XGBoost	**0.816**	**0.445**

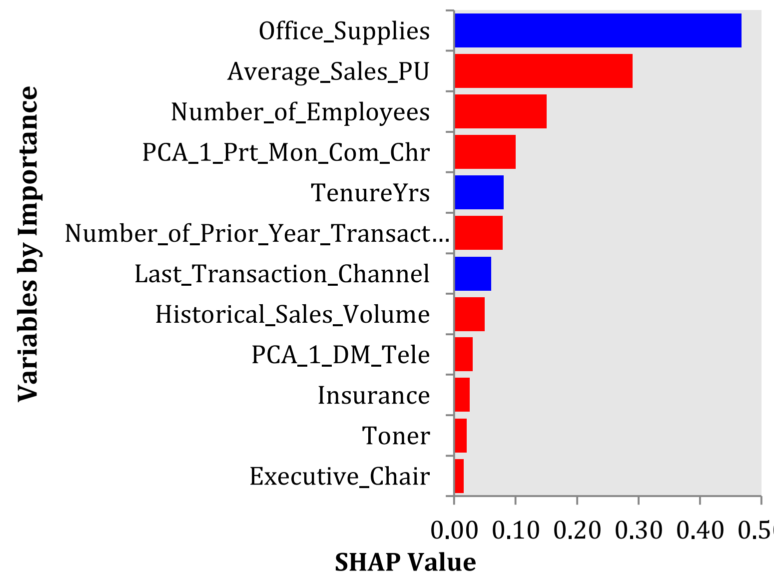
WHAT DOES IT MEAN?

XGBoost has the lowest RMSE, which indicates the absolute fit of the model to actual data - the lower the RMSE value, the better the model.

A few key variables showed significant when calculating the expected sales value. For example:

- ***Office Supplies*** *– prior purchased of office supplies lower the expected value;*
- ***Average Sales per unit*** *– the higher the value, the higher the expected value;*
- ***TenureYrs*** *– the longer the tenure, the lower the expected value.*

Result from Best Regression Model
Variables Importance Plot



(Red = Positive Impact, Blue = Negative Impact)

Illustration:

Two examples showed how expected values are calculated.

Note: Some features/expected value shown here have been encoded/transformed.

Note1: R2 Score is the coefficient of determination. It measures the quality of fit.
Note2: RMSE (aka Root Mean Square Error) measure the average magnitude of the residuals.

4 Analysis Result – Let's combine them!

A few examples on how these models are applied:

Prediction		Actual	
Propensity Model	Regression Model	Test Results	
Probability (Sale)	Estimated (Transaction Size)	Responded	Campaign Period Sales
0.7534	$2,579.05	Yes	$2,501.50
0.1372	$0.00	No	$0.00
0.0198	$0.00	No	$0.00
0.9111	$248.72	Yes	$225.60
0.9993	$4,838.12	Yes	$5,560.10

The two models, when combined, do project relatively strong predictive power by

(i) **calculating** the probability if the customer will respond

(ii) **estimating** the expected sales value.

Next stage, how to maximise profit using this!

4 Analysis Result – Expected Profit Calculation

Relevant financials components:

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

E(Profit) = 0.22*Prob(Sale)*Est(Transaction Size) - $8.40*Prob(Sale) - $45.65

Take same examples from previous slide and calculate the expected profit using the financial model above.

Model 1	Model 2	Combined
Probability (Sale)	Estimated (Transaction Size)	Expected Profit $
0.7534	$2,579.05	$375.50
0.1372	$0.00	-$46.80
0.0198	$0.00	-$45.82
0.9111	$248.72	-$3.45
0.9993	$4,838.12	$1,009.56

Apply this calculation onto the test dataset.

4 Analysis Result – Gain/Lift Chart

Rank profit & sort into 10 groups

Decile	Number of Customers	Actual Profitability 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 Profit 100k Cust Base ($K)	Cuml Total Profit 100k Cust Base ($K)
1	794	$ 325.72	$ 328.31	$ 258,622	1243%	$ 3,283	$ 3,257	$ 3,283	$ 3,257
2	795	$ 31.57	$ 34.16	$ 25,099	121%	$ 342	$ 316	$ 3,625	$ 3,573
3	795	$ (4.85)	$ (2.26)	$ (3,853)	-19%	$ (23)	$ (48)	$ 3,602	$ 3,524
4	795	$ (54.05)	$ (51.46)	$ (42,970)	-206%	$ (515)	$ (541)	$ 3,088	$ 2,984
5	795	$ (54.05)	$ (51.46)	$ (42,970)	-206%	$ (515)	$ (541)	$ 2,573	$ 2,443
6	795	$ (54.05)	$ (51.46)	$ (42,970)	-206%	$ (515)	$ (541)	$ 2,058	$ 1,903
7	795	$ (54.05)	$ (51.46)	$ (42,970)	-206%	$ (515)	$ (541)	$ 1,544	$ 1,362
8	795	$ (54.05)	$ (51.46)	$ (42,970)	-206%	$ (515)	$ (541)	$ 1,029	$ 822
9	795	$ (54.05)	$ (51.46)	$ (42,970)	-206%	$ (515)	$ (541)	$ 515	$ 281
10	793	$ (54.05)	$ (51.46)	$ (42,862)	-206%	$ (515)	$ (541)	$ -	$ (259)
Total	**7947**	**$ (2.59)**	**$ -**	**$ (20,811)**	**100%**				

- A campaign without appropriate selection process can lead to losses

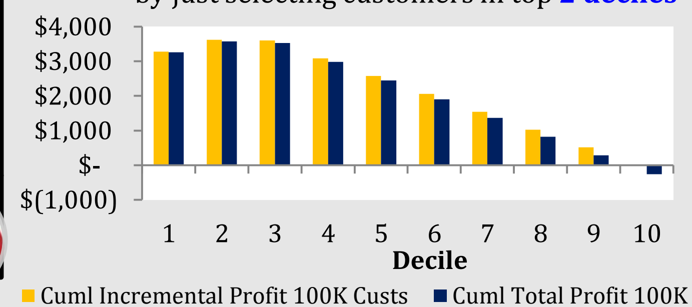
Random Targeting

- Selecting customers based on top decile(s) can maximise company's profit

Selective Targeting

Selective Targeting:

Expected to gain **$3.57M** total profit by just selecting customers in top **2 deciles**



5 Recommendations

1. Test results showed **distinct characteristics** between respondents and non-respondents.

2. Both models have **more than 80% predictive capability** when calculating customer's likelihood to respond and estimating their expected sales values.

3. By adopting **selective targeting** on top deciles, we are able to maximise total profit or allowing the business to **choose the right decile(s)** based on their marketing objective / budget.

4. Propose to deploy using **test-and-control** groups **(i.e. A/B testing)** to gauge model effectiveness and constantly improve the model through adding various touch-points like digital side of things, geo-locations or creating bespoke **customer segmentation** via clustering.

Conclusions

If you interested on how the models are constructed, please find details in the following link:
https://github.com/ucdcsl55/projects

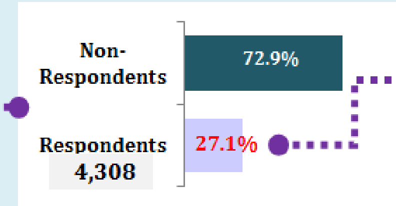
Appendix – Respondents' Characteristics Dashboard

RESPONDENTS' CHARACTERISTICS DASHBOARD

Repurchase Method	Relative Index
AUTO RENEW	117
NOTICE	94

Last Transaction Channel	Relative Index
AUTO RENEW	114
BILLING	123
BRANCH (PHONE)	116
BRANCH (POS)	98
IT	96
MAIL	96
PHONE	88
WEB	99

Language	Relative Index
Arabic	103
Chinese	54
English	102
French	44
German	95
Greek	123
Hebrew	38
Hindi	36
Italian	54
Japanese	134
Korean	0
Pashto	90
Polish	54
Portuguese	87
Russian	67
Spanish	72
Thai	269
Vietnamese	77



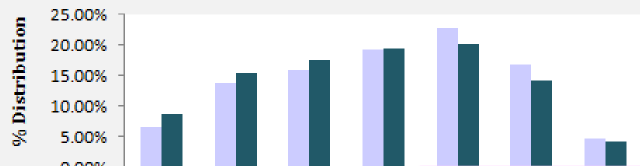
Historical Sales Band	Relative Index
>= $1.5M	76
$800k- $1.5M	88
$500K- $800K	90
$300K- $500K	99
$150K- $300K	113
$ 50K - $150K	119
< $50K	111

Prior Yr Trans Band	Relative Index
>=40	451
30-39	563
20-29	120
15-19	117
5-14	82
<5	53

Tenure Band	Relative Index
>=50 Yrs	1539
40-49 Yrs	985
30-39 Yrs	47
25-29 Yrs	172
20-24 Yrs	8
10-19 Yrs	79
5-9 Yrs	73
<5 Yrs	101

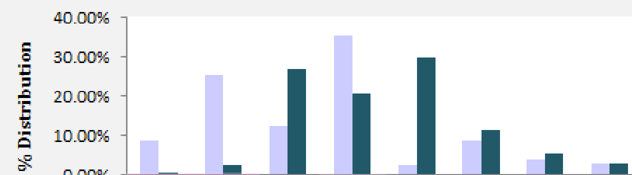
Num of Employees	Relative Index
1-5	112
6-10	93
11-50	114
51-100	61
101-500	115
500+	133

Historical Sales Volume by Bucket
Respondents: Relatively lower historical sales, <= $150K



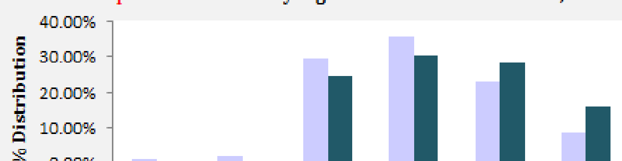
	>= $1.5M	$800k- $1.5M	$500K- $800K	$300K- $500K	$150K- $300K	$ 50K - $150K	< $50K
Respondents	6.62%	13.67%	15.92%	19.34%	22.84%	16.95%	4.67%
Non-Respondents	8.69%	15.51%	17.64%	19.47%	20.24%	14.23%	4.20%

Tenure by Bucket
Respondents: Relatively longer tenure, >=25 years



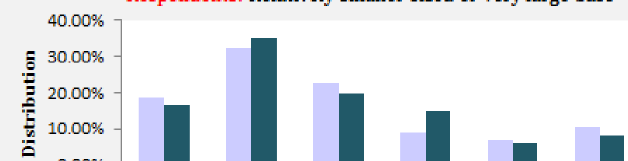
	>=50 Yrs	40-49 Yrs	30-39 Yrs	25-29 Yrs	20-24 Yrs	10-19 Yrs	5-9 Yrs	<5 Yrs
Respondents	8.64%	25.26%	12.44%	35.38%	2.44%	8.96%	4.11%	2.79%
Non-Respondents	0.56%	2.56%	26.73%	20.58%	29.81%	11.35%	5.65%	2.74%

Number of Prior Year Transactions by Bucket
Respondents: Relatively higher number of transactions, >= 15



	>=40	30-39	20-29	15-19	5-14	<5
Respondents	1.09%	2.04%	29.46%	35.63%	23.14%	8.64%
Non-Respondents	0.24%	0.36%	24.60%	30.37%	28.29%	16.14%

Number of Employees by Bucket
Respondents: Relatively smaller sized or very large base

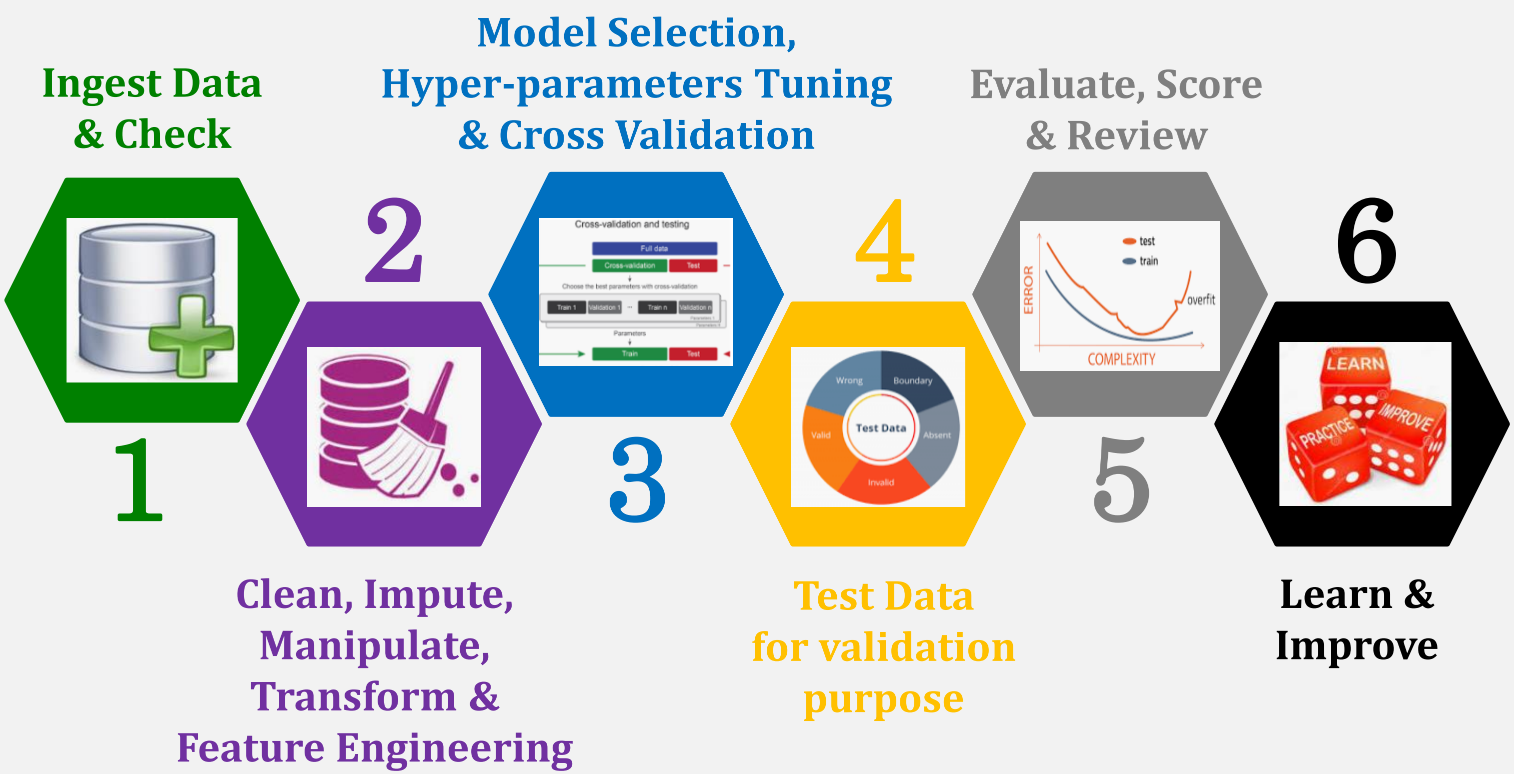


	1-5	6-10	11-50	51-100	101-500	500+
Respondents	18.71%	32.40%	22.52%	8.94%	6.82%	10.61%
Non-Respondents	16.64%	34.96%	19.74%	14.73%	5.93%	7.99%

A2

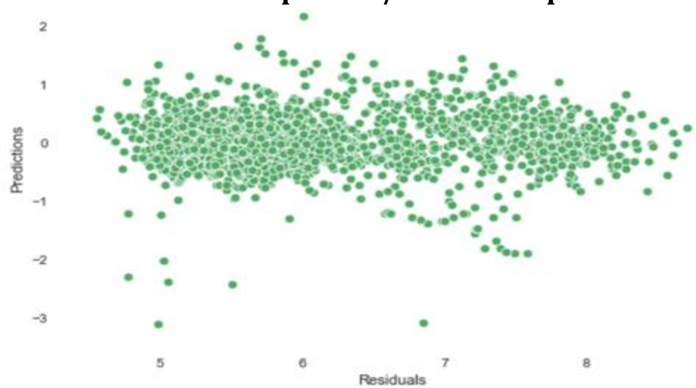
Appendix – Machine Learning Cookbook

Machine Learning Process

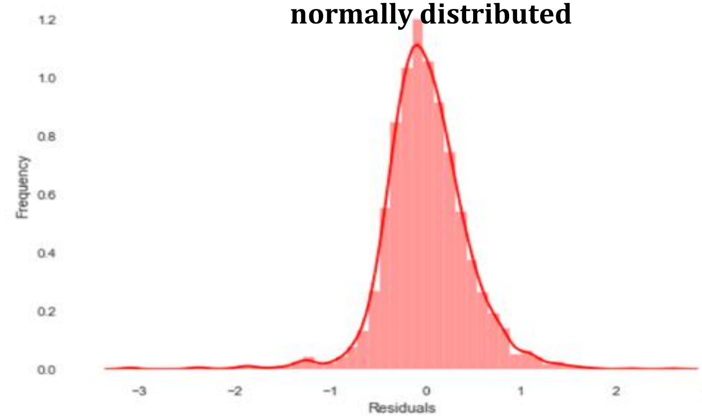


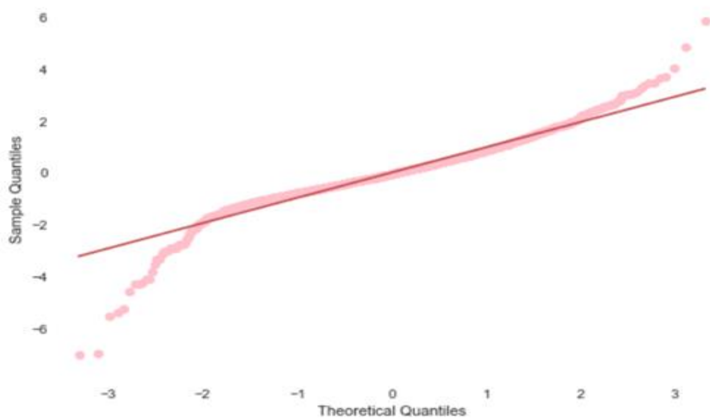
A3 Appendix – Model Diagnostics Review

Residuals are scattered randomly.
Ensure no pattern/relationship exhibit

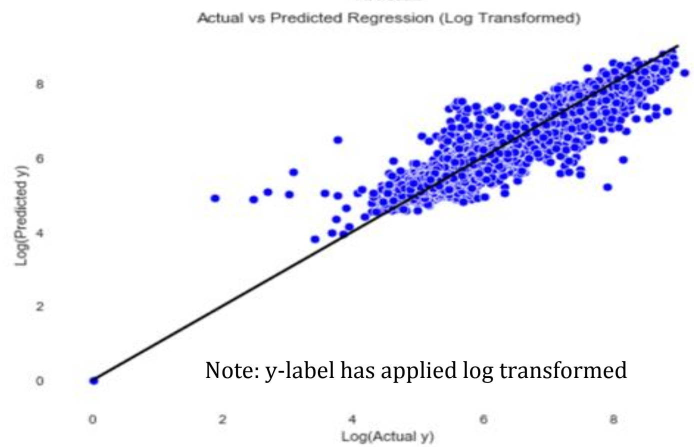


Residuals are expected to be normally distributed





Q-Q plot lies on a straight line except some outliers in lower and upper ends but insignificant



Predicted values plotted along straight line (assess how close it predicts against the actual value)

Officework Supplies
Your one-stop center