Causal Analytics “Econometric model in NPS”

TEAM MEMBERS

Dania Crivelli A20071614
Rhoda Alawiye A20392590

Table of contents

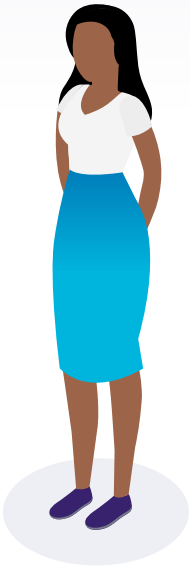
Our team

Dania Crivelli

1st year MsBanDS student

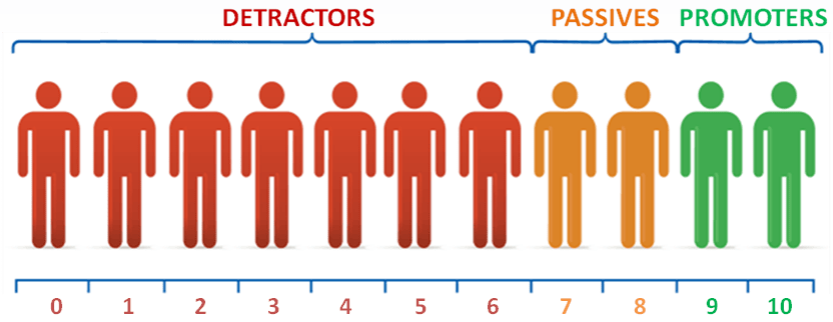
Rhoda Alawiye

1st year MsBanDS student



A. Business Understanding

- What happens to the probability to promote when a promoter topic is increased by 10% and a detractor topic is reduced by 10%?



NPS = **% of Promoters** (9s and 10s) − **% of Detractors** (0s through 6s)

CRISP-DM MODEL

We followed the CRISP-DM Model, which is a methodology followed for the data mining processes.

DATA DICTIONARY

Alphabetic List of Variables and Attributes

#	Variable	Type	Len	Format	Label
1	Account_Class	Char	22		Account_Class
3	Agent_Cont_DateTime	Char	20		Agent_Cont_DateTime
2	Agent_Contact_Date	Num	8	MMDDYY10.	Agent_Contact_Date
4	Agent_Country	Char	2		Agent_Country
5	Agent_Department	Char	37		Agent_Department
6	Agent_ID	Char	21		Agent_ID
7	Agent_Location	Char	43		Agent_Location
8	Agent_Tenure	Num	8		Agent_Tenure
13	CLV1_Num	Num	8		CLV1_Num
14	Customer_Engage_Status	Char	11		Customer_Engage_Status
15	Customer_ID	Num	8		Customer_ID
16	Customer_Tenure	Num	8		Customer_Tenure
17	EM_CLASSIFICATION	Char	11		
18	EM_EVENTPROBABILITY	Num	8		
19	EM_PROBABILITY	Num	8		
20	General_Comments	Char	10000		General_Comments
21	INVSQR_CLV1_Num	Num	8		Transformed CLV1_Num
22	I_NPS_Detract_Pass_Prom_Text	Char	32		Into: NPS_Detract_Pass_Prom_Text
23	LG10_MBR_NTPV_12_Mth_S	Num	8		Transformed MBR_NTPV_12_Mth_S

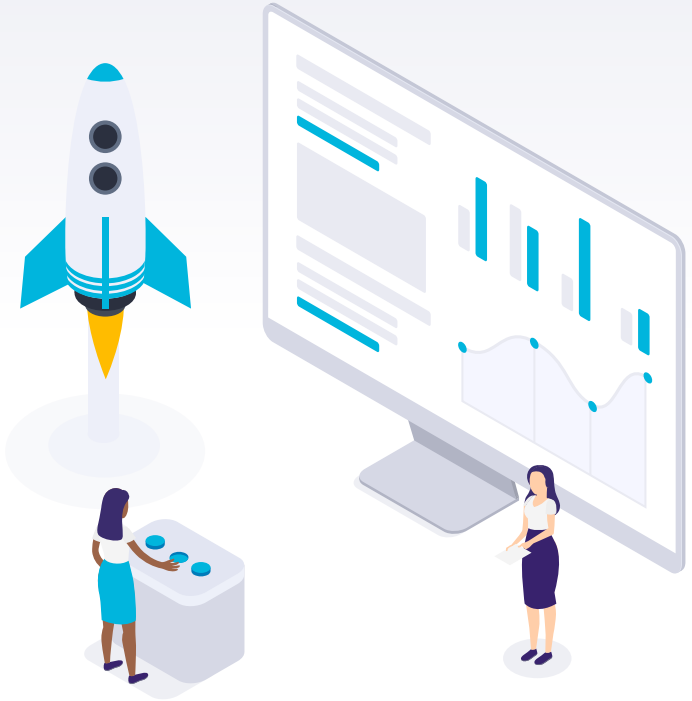
Data summary:

- 72900 observations
- Total 42 variables
- 12 string variables
- 30 integer variables

1

Methodology

Steps followed in our analysis



B. Data Cleaning

- The first step we took in the analysis was to first subset the data to only include “Active” customers in the “Customer_engage_status” variable as well as removing any “Guest-Upgradable” observations in the “Account_Class” variable.

```
[1]: import pandas as pd

[3]: df = pd.read_csv("casuser/OSUM3_PREDICTNPS.csv")

[4]: df_2 = df[df["Account_Class"] != 'Guest-Upgradeable']

[5]: df_3 = df_2[df_2["Customer_Engage_Status"] == 'Active']
     df_3.to_csv('updated1_df.csv')
```

- The format for the est_spend variable was changed because it has a dollar sign which made it appear as a character variable.

est_spend
$7
$192
$6
$428
$6,559
$300
$367
$765

B2. Data Cleaning

- This is the code we used to change the format of the string to integer for Est_spend variable

```
cas mySession sessopts=(caslib=casuser timeout=2400 locale="en_US" metrics='true');
caslib _all_ assign;

cas;
caslib _all_ assign;

ods graphics on;
title 'Using SAS9 ETS Proc Model with SAS Studio in Viya';

data work.import_new;
     set work.import;
est_spend_num = input(est_spend,dollar9.);
run;
```

C. Data Preparation

```
proc summary data= Work.import_new;
class Week;
var Agent_Tenure Customer_Tenure est_spend_num category_1 category_2 category_3 category_4 P_NPS_Detract_Pass_ProPROM(
output out = work.import_new2 (drop= Agent_Tenure Customer_Tenure est_spend_num category_1 category_2 category_3 categ
 mean = mean(Agent_Tenure)= avg_agent_tenure
  mean = mean(Customer_Tenure)= avg_customer_tenure
  mean = mean(est_spend_num)= avg_est_spend_num
  sum =  sum(category_1) = sum_category_1
  sum =  sum(category_2) = sum_category_2
  sum =  sum(category_3) = sum_category_3
  sum =  sum(category_4) = sum_category_4;
run;

ods graphics on;
```

- **Proc Summary** was used in the analysis to aggregate the data by weeks and compute summary statistics for the variables of interest.
- The "var" statement specifies the summarized variables.
- The mean statement was used to compute the average values of each **interval** variables and the sum statement computes the sum of each of the four **categorical** variable.

Overall, these steps helps to provide a more manageable and condensed data set for the analysis while still retaining the necessary information for the 7-weeks data.

avg_customer_tenure	avg_est_spend_num	sum_category_1	sum_category_2
1635.6880759	471.44022648	9643	7399
1637.5955418	470.26937022	1303	1045
1638.4806608	469.06459844	1290	1059
1642.9458255	463.82767483	1305	957
1648.1931742	470.97027063	1409	1030
1629.1729072	457.84050973	1306	1050
1633.0705427	470.72969729	1360	964
1623.9238537	491.42025684	1670	1294

C2. Regression Model Specification

```
proc model data=work.import_new2 plots=all;
var b0 b1 b2 b3 b4 b5 b6 b7;
label b0 = 'Intercept'
      b1 = 'Avg_AgentTenure'
      b2 = 'Avg_CustomerTenure'
      b3 = 'Avg_est_spend'
      b4 = 'Sum_cat1'
      b5 = 'Sum_cat2'
      b6 = 'Sum_cat3'
      b7 = 'Sum_cat4';
P_NPS_Detract_Pass_ProPROMOTERS = b0 + b1*avg_agent_tenure + b2*avg_customer_tenure + b3*avg_est_spend_num +
b4*sum_category_1 + b5*sum_category_2 + b6*sum_category_3 + b7*sum_category_4;
fit P_NPS_Detract_Pass_ProPROMOTERS /white breusch=(1 income) out=casuser.econ_output
outpredict outactual details;
```

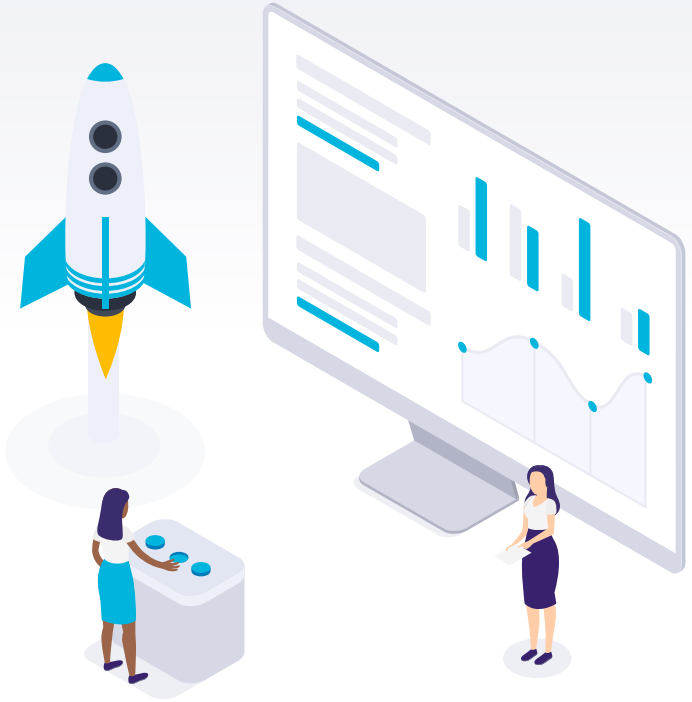
- Proc Model is used to specify the linear regression model to predict the probability of being a promoter based on the variables of interest.
- The “var” statement specifies the variables to be included in the model which are the intercepts with the interval and categorical variables.
- The “label” statement is used for the purpose of easier interpretation of results.
- The regression equation is used to specify how the predictor variables are combined to predict the target variable(P_NPS_Detract_Pass_ProPROMOTERS).
- This helps to understand the relationship between the predictor variables and the response variables which will enable us to identify the variables that are of the most importance in predicting the target/response variable.

D. Modeling

```
weight _freq_;
estimate 'Safe/Easy -10%, Prob+10%' b0 + b1*(597.42)*1.1 + b2*(1635.69)*1.1 + b3*(471.44)*0.9 + b4*(9643)*0.9 +
            b5*(7399)*0.9 + b6*(18006)*1.1 + b7*(10624)*0.9;
estimate 'Safe/Easy no change' b0 + b1*(597.42) + b2*(1635.69) + b3*(471.44) + b4*(9643) +
            b5*(7399) + b6*(18006) + b7*(10624);
title 'Proc Model for Simple Linear Combinations';
run;
title;
ods graphics off;
```

- The estimate statement is used to estimate the predicted value of the target variable.
- The first “estimate” statement tests what happens to the probability to promote when a promoter topic is increased by 10% while the detractor topic is reduced by 10%.
- The second “estimate” statements tests what happens to the probability to promote when no changes are made to the variables in the model.
- These are to help understand how the changes in the independent variables affect the predicted variable.

Analysis Results



Analysis Result 1

Proc Model for Simple Linear Combinations

The MODEL Procedure

Model Summary	
Model Variables	1
Parameters	8
ID Variables	1
Equations	1
Number of Statements	5

Model Variables	P_NPS_Detract_Pass_ProPROMOTERS
Parameters	b0 b1 b2 b3 b4 b5 b6 b7
Equations	P_NPS_Detract_Pass_ProPROMOTERS

The Equation to Estimate is	
P_NPS_Detract_Pass_ProPROMOTERS =	F(b0(1), b1(avg_agent_tenure), b2(avg_customer_tenure), b3(avg_est_spend_num), b4(sum_category_1), b5(sum_category_2), b6(sum_category_3), b7(sum_category_4))

Observations will be weighted by	_FREQ_

Storage Requirements for this Problem	
Order of XPX Matrix	9
Order of Cross Matrix	9
Total Nonzero Derivatives	8
Distinct Variable Derivatives	7
Size of Cross matrix	9024

NOTE: At OLS Iteration 1 convergence assumed because OBJECTIVE=1.510176E-20 is almost zero (<1E-12).

- The linear combination in which each independent variable is multiplied by its respective parameter and then added to the intercept parameter.
- The OLS(Ordinary Least Squared) algorithm shores that the objective function converges at the first iteration because it was very close to zero.

Analysis Result 2

Nonlinear OLS Estimates					
Term	Estimate	Approx Std Err	t Value	Approx Pr > |t|	Label
Safe/Easy-10%, Prob+10%	1.315512	0	.	.	b0 + b1*(597.42)*1.1 + b2*(1635.69)*1.1 + b3*(471.44)*0.9 + b4*(9643)*0.9 + b5*(7399)*0.9 + b6*(18006)*1.1 + b7*(10624)*0.9

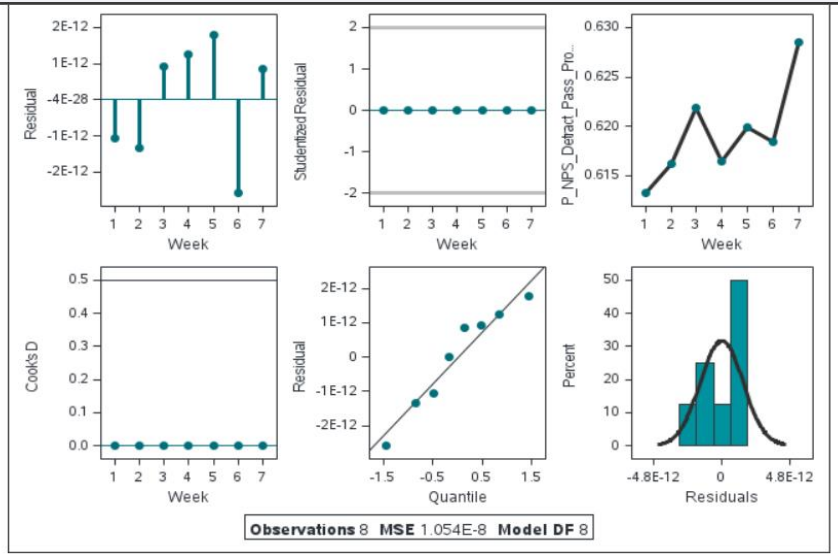
Nonlinear OLS Estimates					
Term	Estimate	Approx Std Err	t Value	Approx Pr > |t|	Label
Safe/Easy no change	0.619599	0	.	.	b0 + b1*(597.42) + b2*(1635.69) + b3*(471.44) + b4*(9643) + b5*(7399) + b6*(18006) + b7*(10624)

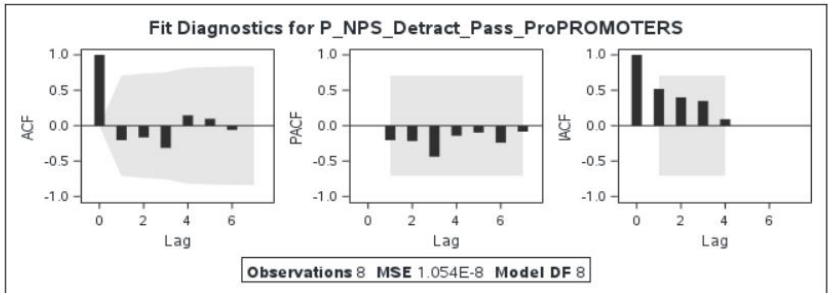
Number of Observations		Statistics for System	
Used	8	Objective	1.51E-20
Missing	0	Objective*N	1.208E-19
Sum of Weights	108956		

Heteroscedasticity Test					
Equation	Test	Statistic	DF	Pr > ChiSq	Variables
P_NPS_Detract_Pass_ProPROMOTERS	White's Test	8.00	7	0.3326	Cross of all vars
	Breusch-Pagan	.	0	.	1, income

- The first “estimate” statement that tests what happens to the probability to promote when a promoter topic is increased by 10% while the detractor topic is reduced by 10% is 1.315512.
- The second “estimate” statements that tests for what happens to the probability to promote when no changes are made to the variables in the model is 0.619599.
- The White’s test had a value of 0.3326 for the Pr > ChiSq at 7 DF across all variables and the Breusch-pagan test was null for the test of heteroscedasticity.
- This suggests that there is no significant evidence of heteroscedasticity in the model.

Analysis Result 3





- The fit diagnostics report for the target variable P_NPS_Detract_Pass_ProPROMOTERS showing the Residual plot, Studentized residual, Cook's D and the quantile plot which are used to test for the linearity, normality and heteroscedasticity of errors of the regression model.

- The fit diagnostics report for the target variable P_NPS_Detract_Pass_ProPROMOTERS also show the ACF(Autocorrelation function) which is used to test for the autocorrelation, PACF(Partial Autocorrelation function) used to measure the correlation between a time series and its lagged values after removing the effects of intermediate lags and WCF(Wavelet Coherent Function) used to measure the correlation between two time series across different frequencies.

Analysis Result 4

Obs	Week	_ESTYPE_	_TYPE_	_WEIGHT_	_FREQ_	P_NPS_Detract_Pass_ProPROMOTERS	avg_agent_tenure	avg_customer_tenure	avg_est_spend_num	sum_category_1	sum_category_2	sum_catego
1	.	OLS	ACTUAL	54478	54478	0.6196013779	597.42325342	1635.6880759	471.440	9643	7399	1
2	.	OLS	PREDICT	54478	54478	0.6196013779	597.42325342	1635.6880759	471.440	9643	7399	1
3	1	OLS	ACTUAL	7447	7447	0.6131998176	595.84141265	1637.5955418	470.269	1303	1045	
4	1	OLS	PREDICT	7447	7447	0.6131998176	595.84141265	1637.5955418	470.269	1303	1045	
5	2	OLS	ACTUAL	7446	7446	0.6161447965	603.97743755	1638.4806608	469.065	1290	1059	
6	2	OLS	PREDICT	7446	7446	0.6161447965	603.97743755	1638.4806608	469.065	1290	1059	
7	3	OLS	ACTUAL	7402	7402	0.6218592049	587.54039449	1642.9458255	463.828	1305	957	
8	3	OLS	PREDICT	7402	7402	0.6218592049	587.54039449	1642.9458255	463.828	1305	957	
9	4	OLS	ACTUAL	7501	7501	0.6164521589	607.53112918	1648.1931742	470.970	1409	1030	
10	4	OLS	PREDICT	7501	7501	0.6164521589	607.53112918	1648.1931742	470.970	1409	1030	
11	5	OLS	ACTUAL	7478	7478	0.6198843291	597.5062851	1629.1729072	457.841	1306	1050	
12	5	OLS	PREDICT	7478	7478	0.6198843291	597.5062851	1629.1729072	457.841	1306	1050	
13	6	OLS	ACTUAL	7499	7499	0.6184210631	591.73223096	1633.0705427	470.730	1360	964	
14	6	OLS	PREDICT	7499	7499	0.6184210631	591.73223096	1633.0705427	470.730	1360	964	
15	7	OLS	ACTUAL	9705	9705	0.6285715226	597.66718187	1623.9238537	491.420	1670	1294	
16	7	OLS	PREDICT	9705	9705	0.6285715226	597.66718187	1623.9238537	491.420	1670	1294	

- This shows the OLS results of the actual and predicted values for each of the variables of the regression model for the 7-weeks.

Conclusion

Safe/Easy -10%/Prob+10%

The probability to promote when changes are made to the promoter topic and the detractor topic is 31.55%.

Safe/Easy no change

The probability to promote when no changes are made to the variables in the model is 6.19%


THANK YOU !
BEGIN.