Lapse prediction in Life insurance contract

Logistic Regression With SPSS

Preventing Policy Lapse Proactively

Persistency is a key driver for successful insurance businesses. We just cannot let existing customers churn and then terminate their policies. Data science proactively alerts, and actions are necessary to address policy lapse.

The typical solution approach is to devise a logistic regression model to predict the likelihood of a lapse of policies. There are several data points that go in as inputs to this model, such as:

- Customer Demographics – Gender, Age, Race, Income, Nationality, Marital Status
- Customer Interaction mode and frequency with company – Email, Phone, others (fax, letters)
- Number and type of insurance products customers have purchased from the company
- Policy details – Agent, Sum Insured, Premium, Term
- Each event for the policy – Inception, Lapse, Claim, Reinstatement, Cancel, Surrender, Mature

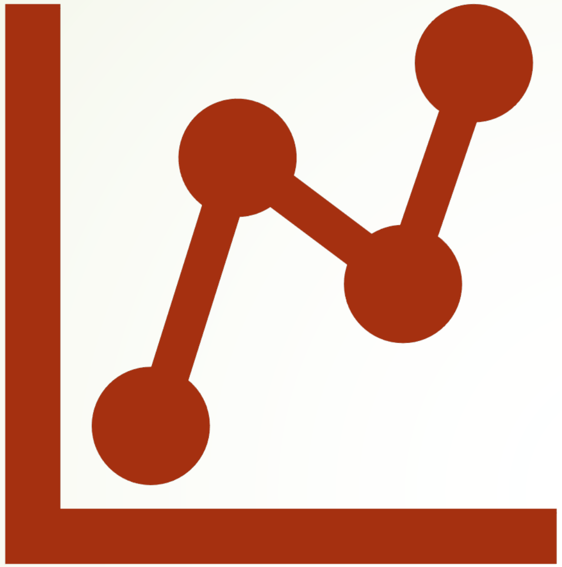
The model output helps in predicting whether a certain customer profile is likely to lapse or not. It also provides indicators on significant factors impacting lapse, for example, Age, Premium level ,Channel of distribution, Customer interaction etc. that can help you take focused actions.

Data Description

The sample data for analysis has a total of 1340 policies (434 Lapse, 907 inforce) with sample features :

- ***Lapse:*** 0 = Policy In-force, 1 = Lapsed
- ***NumOfReinstated:*** Number of reinstated
- ***NumOfClaims:*** Number of Claims
- ***NumOfEmails:*** Frequency of contact by Email
- ***NumOfCalls:*** Frequency of contact by Phone call
- ***PO Sex:*** PO Sex, Male or Female
- ***PO Age:*** PO Age in years
- ***PO_Married:*** PO Marial status

- ***INS_Age:*** Insured Age in years
- ***INS_Sex:*** Insured Sex, Male or Female
- ***Occupation:*** PO Occupation Classes (with 4 classes)
- **Premium:** Premium fees
- ***Coverage Period***: 1-5 years, 5-10 years, 10-20+ years
- ***PaymentTerm:*** Preimum Payment Term: Monthly, Quarterly, Semi-annual, Annually
- ***DistributionChannel:*** Company Agent, Bancasurance, Corp Channel, General Agent, Others.
- ***AgentYearSVR***: Years of Experience of servicing agent



Exploratory Data Analysis (EDA)

Summary

Overal checking data set, we found:

- 32% policies is lapsed.
- 53% of Policy Owner are male, similarly, 52% of INS person are male.
- 14% PO is also Insured.
- PO ages range from 22 to 59.
- We also detected feature NumberOfClaims, NumberOfEmails, NumberOfCalls and Phone_registered with missing values.

Case Processing Summary

	Cases					
	Included		Excluded		Total	
	N	Percent	N	Percent	N	Percent
Lapsed	1341	100.0%	0	0.0%	1341	100.0%
NumOfReinstated	1341	100.0%	0	0.0%	1341	100.0%
NumOfClaims	1339	99.9%	2	0.1%	1341	100.0%
NumOfEmails	1340	99.9%	1	0.1%	1341	100.0%
NumOfCalls	1337	99.7%	4	0.3%	1341	100.0%
Phone_registered	1329	99.1%	12	0.9%	1341	100.0%
PO_Age	1341	100.0%	0	0.0%	1341	100.0%
PO Sex	1341	100.0%	0	0.0%	1341	100.0%
PO_is_INS	1341	100.0%	0	0.0%	1341	100.0%
INS_Age	1341	100.0%	0	0.0%	1341	100.0%
Insured Sex	1341	100.0%	0	0.0%	1341	100.0%
Premium	1341	100.0%	0	0.0%	1341	100.0%
AgentYearSVR	1341	100.0%	0	0.0%	1341	100.0%

Case Summaries

	Lapsed	NumOfReinstated	NumOfClaims	NumOfEmails	NumOfCalls	Phone_registered	PO_Age	PO Sex	PO_is_INS	INS_Age	Insured Sex	Premium	AgentYearSVR
N	1341	1341	1339	1340	1337	1329	1341	1341	1341	1341	1341	1341	1341
Mean	.32	.68	.56	1.29	1.13	.70	43.31	.53	.14	39.89	.52	2640.12	1.96
Std. Deviation	.468	1.080	1.059	1.036	1.128	.457	8.858	.499	.343	13.581	.500	2411.586	.818
Minimum	Inforce	0	0	0	0	No	22	female	No	18	female	224	1
Maximum	Lapse	5	5	5	5	Yes	59	male	Yes	64	male	12754	6

Normal Distribution and Variable Correlation

- Since our data is not a small data set, the normality test is not needed; however, to give it a try, we will conduct the normality test for interval variables.
  - The result shows: *PO_Age, INS_Age* and *Premium* are not normally distributed.
  - Other variables are categorical data, hence they are not from normal distribution.

Tests of Normality

	Kolmogorov-Smirnov[a]			Shapiro-Wilk		
	Statistic	df	Sig.	Statistic	df	Sig.
PO_Age	.073	1341	.000[b]	.959	1341	.000
INS_Age	.064	1341	.000[b]	.955	1341	.000
Premium	.189	1341	.000[b]	.813	1341	.000

a. Lilliefors Significance Correction

b. $p<.05$, reject Ho of Normal Distribution

Normal Distribution and Variable Correlation

- Since our variables do not follow the Gaussian distribution (normally distributed), the nonparametric correlation Spearman's rho was computed instead of the conventional Pearson Coefficient.
  a. *INS_Age* has a possitive correlation (.071) with *PO_is_INS* leading assumption older insured person are PO.
  b. *NumOfReinstated* and *NumOfClaims* hold the positive significant correlation with *NumOfCalls, NumOfEmail*, that is more communication, contact to resolve Client requrest.
  c. *Occupation* has a negative correlation with *INS_Age* (-.146) and *Premium* (-.397) showing that PO with occupation class #1 or #2 pay more premium to their older insured. Positive correlation between *Premium* and *INS_Age* (.514) lead to the same assumption.
  d. We found no correlation between *PO_Sex* or *AgentYearSRV* with other varriables.

Correlations

			PO_Age	PO_is_INS	INS_Age	PO Sex	NumOfReinstated	NumOfClaims	NumOfEmails	NumOfCalls	Occupation	Premium	AgentYearSVR
Spearman's rho	PO_Age	Correlation Coefficient	1.000	-.040	.061*	.046	-.012	.009	.031	.011	.041	-.009	-.044
		Sig. (2-tailed)	.	.148	.024	.093	.652	.728	.260	.696	.134	.754	.109
		N	1341	1341	1341	1341	1341	1339	1340	1337	1341	1341	1341
	PO_is_INS	Correlation Coefficient	-.040	1.000	.071**	.053	.018	.014	.030	-.016	.035	.004	-.005
		Sig. (2-tailed)	.148	.	.009	.051	.512	.615	.270	.549	.202	.876	.856
		N	1341	1341	1341	1341	1341	1339	1340	1337	1341	1341	1341
	INS_Age	Correlation Coefficient	.061*	.071**	1.000	.003	-.037	-.025	-.024	-.001	-.146**	.514**	.006
		Sig. (2-tailed)	.024	.009	.	.899	.179	.361	.384	.969	.000	.000	.814
		N	1341	1341	1341	1341	1341	1339	1340	1337	1341	1341	1341
	PO Sex	Correlation Coefficient	.046	.053	.003	1.000	-.019	.017	-.020	.019	.014	-.041	-.005
		Sig. (2-tailed)	.093	.051	.899	.	.491	.534	.456	.489	.599	.136	.866
		N	1341	1341	1341	1341	1341	1339	1340	1337	1341	1341	1341
	NumOfReinstated	Correlation Coefficient	-.012	.018	-.037	-.019	1.000	-.013	.234**	.237**	-.027	.041	-.002
		Sig. (2-tailed)	.652	.512	.179	.491	.	.635	.000	.000	.330	.135	.941
		N	1341	1341	1341	1341	1341	1339	1340	1337	1341	1341	1341
	NumOfClaims	Correlation Coefficient	.009	.014	-.025	.017	-.013	1.000	.166**	.280**	.016	-.061*	.025
		Sig. (2-tailed)	.728	.615	.361	.534	.635	.	.000	.000	.568	.025	.358
		N	1339	1339	1339	1339	1339	1339	1338	1335	1339	1339	1339
	NumOfEmails	Correlation Coefficient	.031	.030	-.024	-.020	.234**	.166**	1.000	.224**	-.037	-.001	-.006
		Sig. (2-tailed)	.260	.270	.384	.456	.000	.000	.	.000	.174	.964	.827
		N	1340	1340	1340	1340	1340	1338	1340	1336	1340	1340	1340
	NumOfCalls	Correlation Coefficient	.011	-.016	-.001	.019	.237**	.280**	.224**	1.000	.002	.001	-.002
		Sig. (2-tailed)	.696	.549	.969	.489	.000	.000	.000	.	.928	.968	.953
		N	1337	1337	1337	1337	1337	1335	1336	1337	1337	1337	1337
	Occupation	Correlation Coefficient	.041	.035	-.146**	.014	-.027	.016	-.037	.002	1.000	-.397**	.014
		Sig. (2-tailed)	.134	.202	.000	.599	.330	.568	.174	.928	.	.000	.607
		N	1341	1341	1341	1341	1341	1339	1340	1337	1341	1341	1341
	Premium	Correlation Coefficient	-.009	.004	.514**	-.041	.041	-.061*	-.001	.001	-.397**	1.000	-.024
		Sig. (2-tailed)	.754	.876	.000	.136	.135	.025	.964	.968	.000	.	.390
		N	1341	1341	1341	1341	1341	1339	1340	1337	1341	1341	1341
	AgentYearSVR	Correlation Coefficient	-.044	-.005	.006	-.005	-.002	.025	-.006	-.002	.014	-.024	1.000
		Sig. (2-tailed)	.109	.856	.814	.866	.941	.358	.827	.953	.607	.390	.
		N	1341	1341	1341	1341	1341	1339	1340	1337	1341	1341	1341

*. Correlation is significant at the 0.05 level (2-tailed).

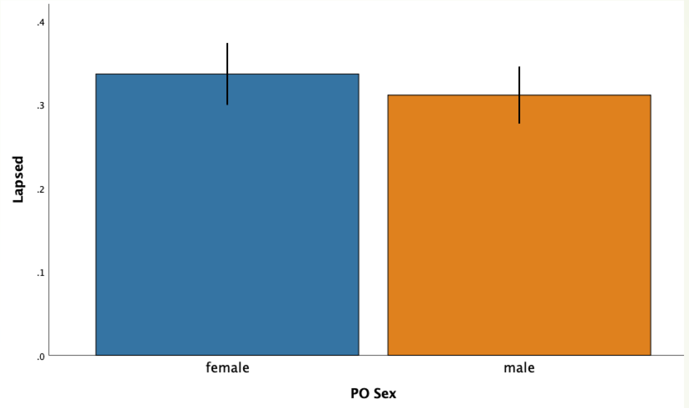
**. Correlation is significant at the 0.01 level (2-tailed).

Features analysis

Gender

Is the likelihood of lapse dependent on gender?

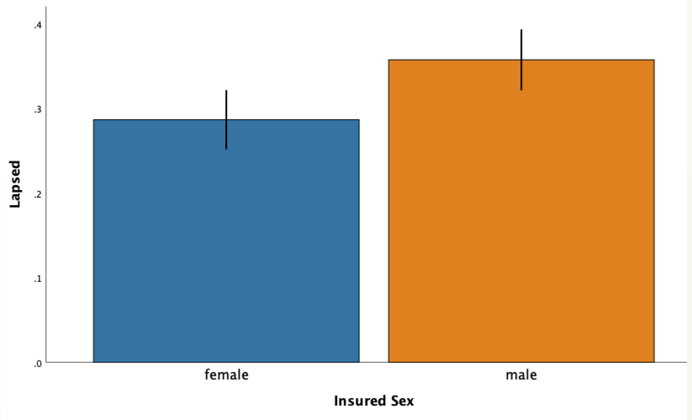
- Policies with female PO are most likely to lapse with a ratio of 34%, while male are lower with a ratio of 31%.
- However, the ratio by Insured person are 36% for male and 29% for female.
- Obviously, Gender is not an important feature to predict lapse.



Lapsed * PO Sex Crosstabulation

% within PO Sex

		PO Sex		
		female	male	Total
Lapsed	Inforce	66.3%	68.8%	67.6%
	Lapse	33.7%	31.2%	32.4%
Total		100.0%	100.0%	100.0%



Lapsed * Insured Sex Crosstabulation

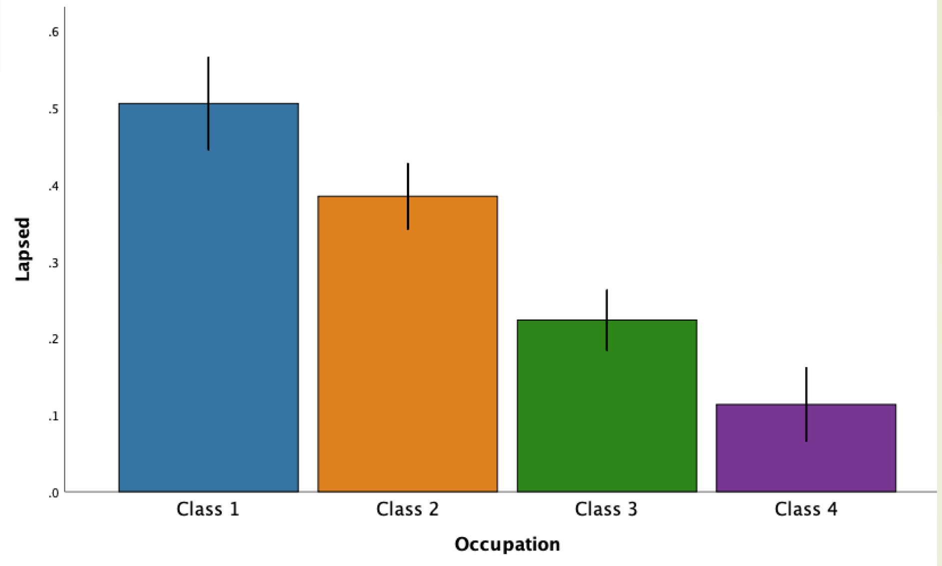
% within Insured Sex

		Insured Sex		
		female	male	Total
Lapsed	Inforce	71.3%	64.2%	67.6%
	Lapse	28.7%	35.8%	32.4%
Total		100.0%	100.0%	100.0%

Occupation Class

Could it be that the Occupation of Policy Owner correlates with the probability of lapse?

- Clients in the occupation class 1 are more likely to lapse than class 2, 3 and 4.
- Occupation is one of the good features for prediction of policy lapse.



Lapsed * Occupation Crosstabulation

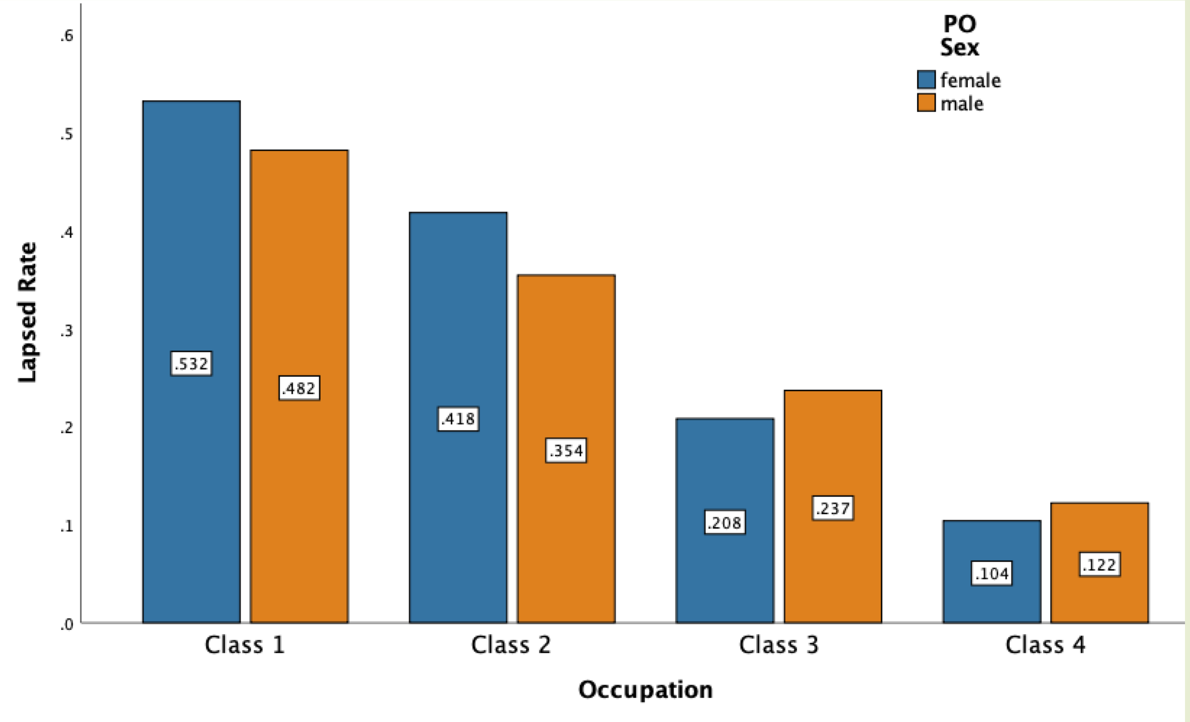
% within Occupation

		Occupation				
		Class 1	Class 2	Class 3	Class 4	Total
Lapsed	Inforce	49.4%	61.5%	77.6%	88.6%	67.6%
	Lapse	50.6%	38.5%	22.4%	11.4%	32.4%
Total		100.0%	100.0%	100.0%	100.0%	100.0%

Occupation and Gender

Does the higher lapse rate in Class 1 have any correlation with gender distribution in which male clients dominate?

- Female in classes 1 and 2 have a higher possibility of lapse than male, but lower in classes 3 and 4.

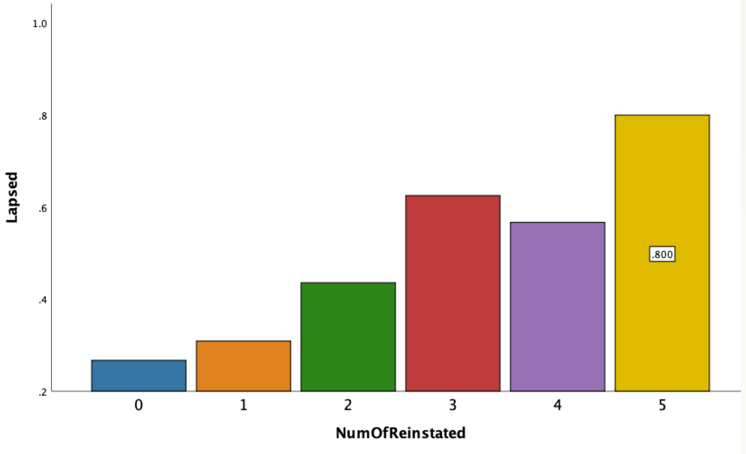


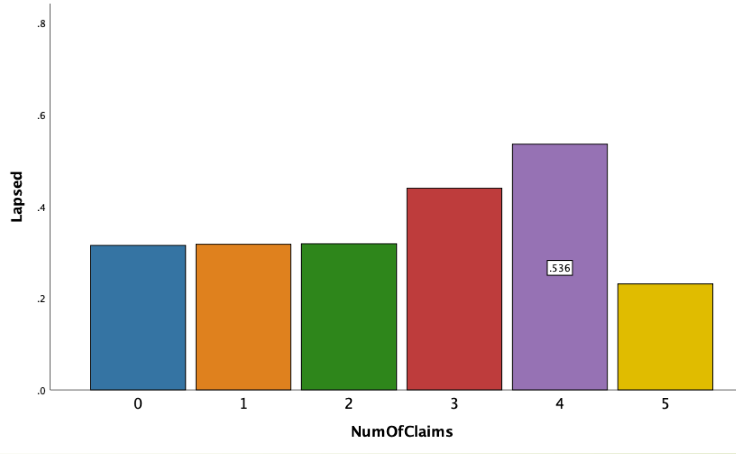
Reinstated and Claims versus Number of Calls and Emails

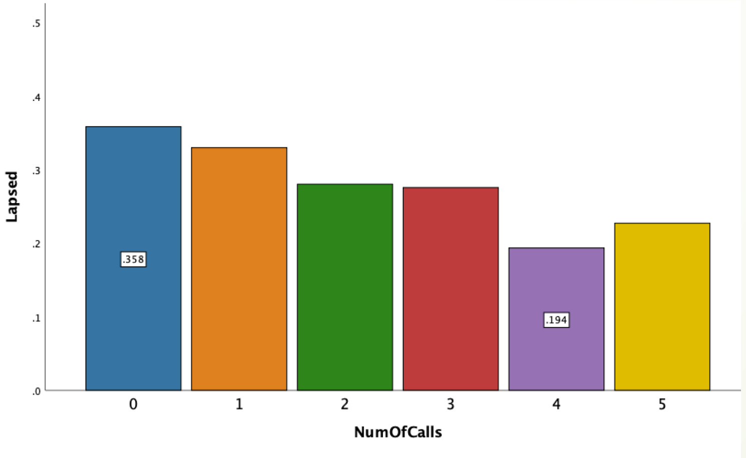
How is the customer interaction and policy events impact to the likelihood of lapse ?

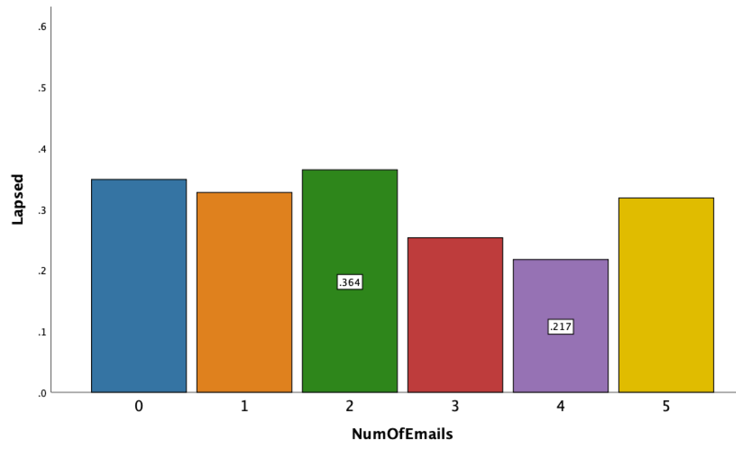
- If the policy have 5 times of reinstated or/and 4 claims, they are likely to lapse.
- If the policy with more contactable by Calls or Email, the possibility of lapse might expect to reduce.

Then we may consider to explore the client interaction (type, frequency) with the policy events.





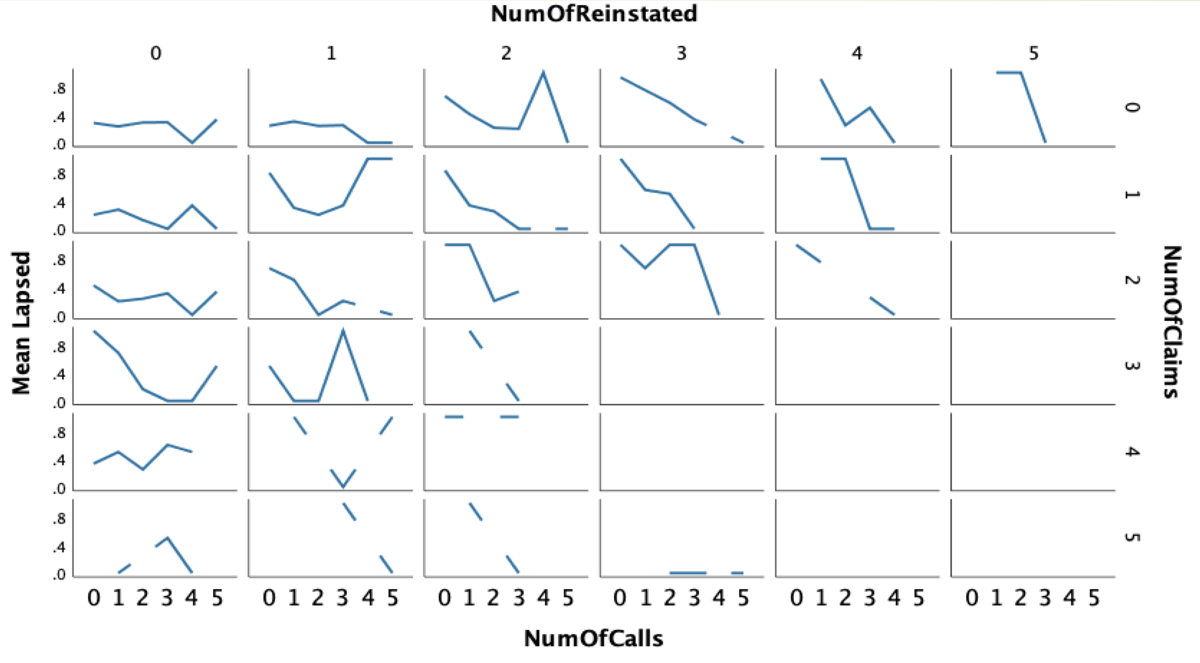


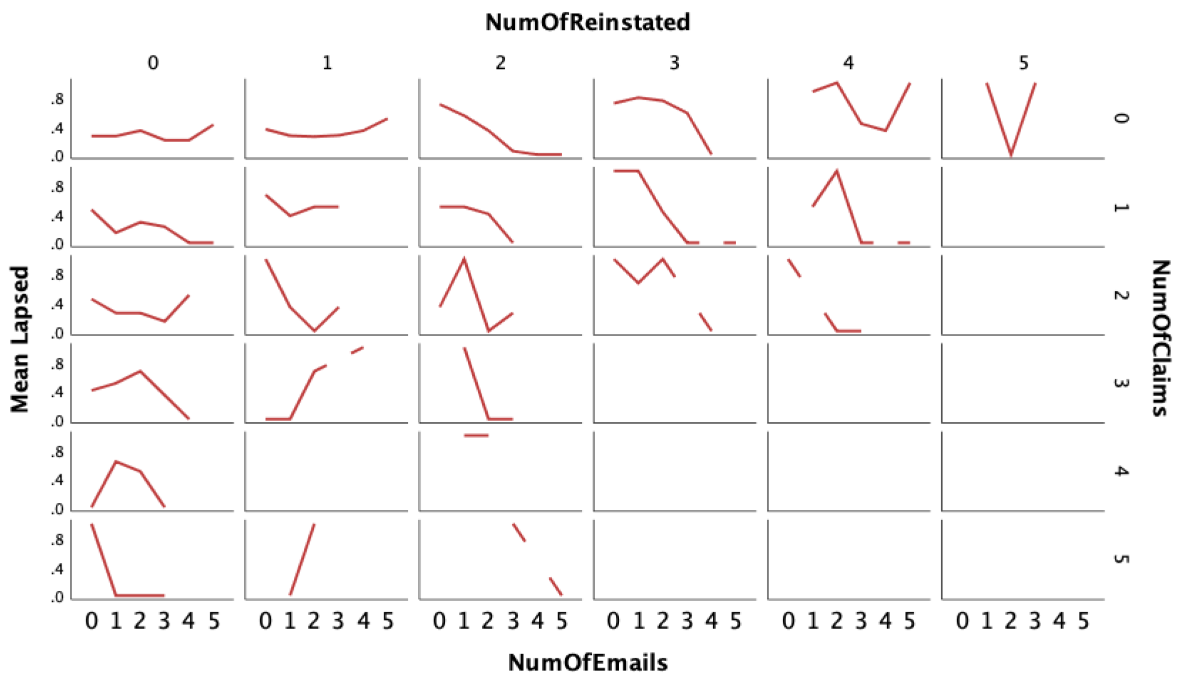


Client Interaction versus Policy events

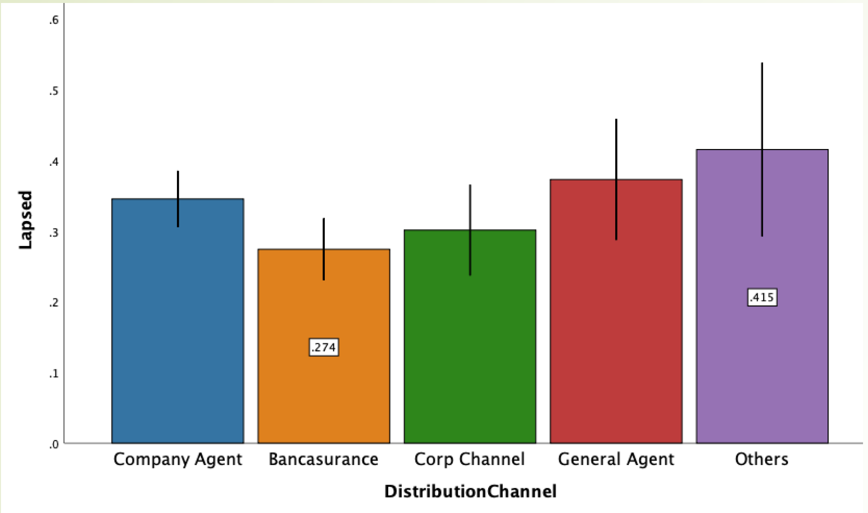
- For policy with less than 2 events (reinstated or/and claims), the lapse rate is not fluctuating change.
- The lapse rate likely to reduce if more interaction for policy have 2 or more events.

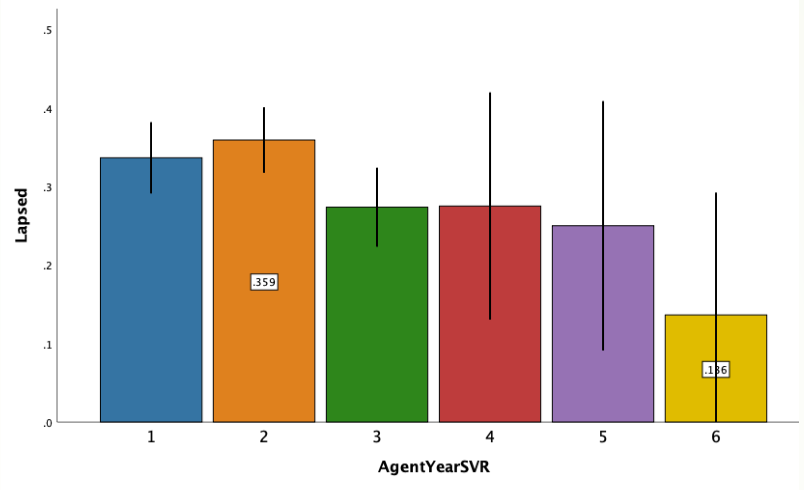
Combination features by Client interaction and Policy events will be considered in features selection.

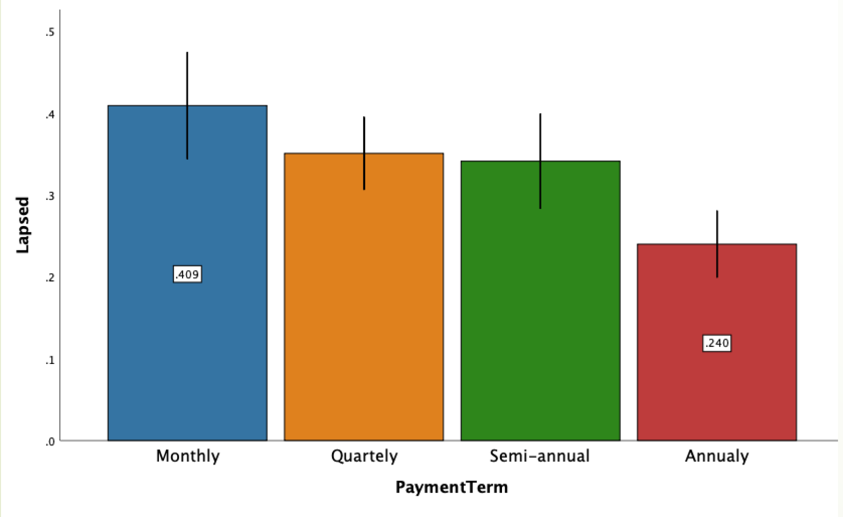


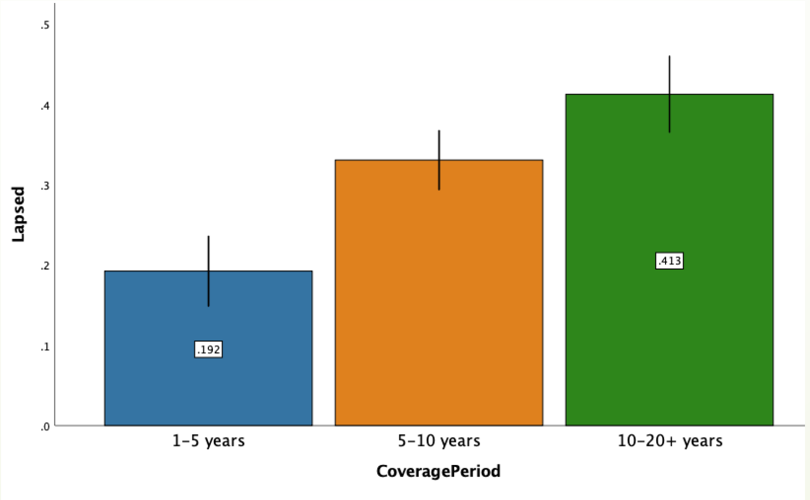


Payment Term, Distribution Channel, AgentYearSVR and Coverage Period





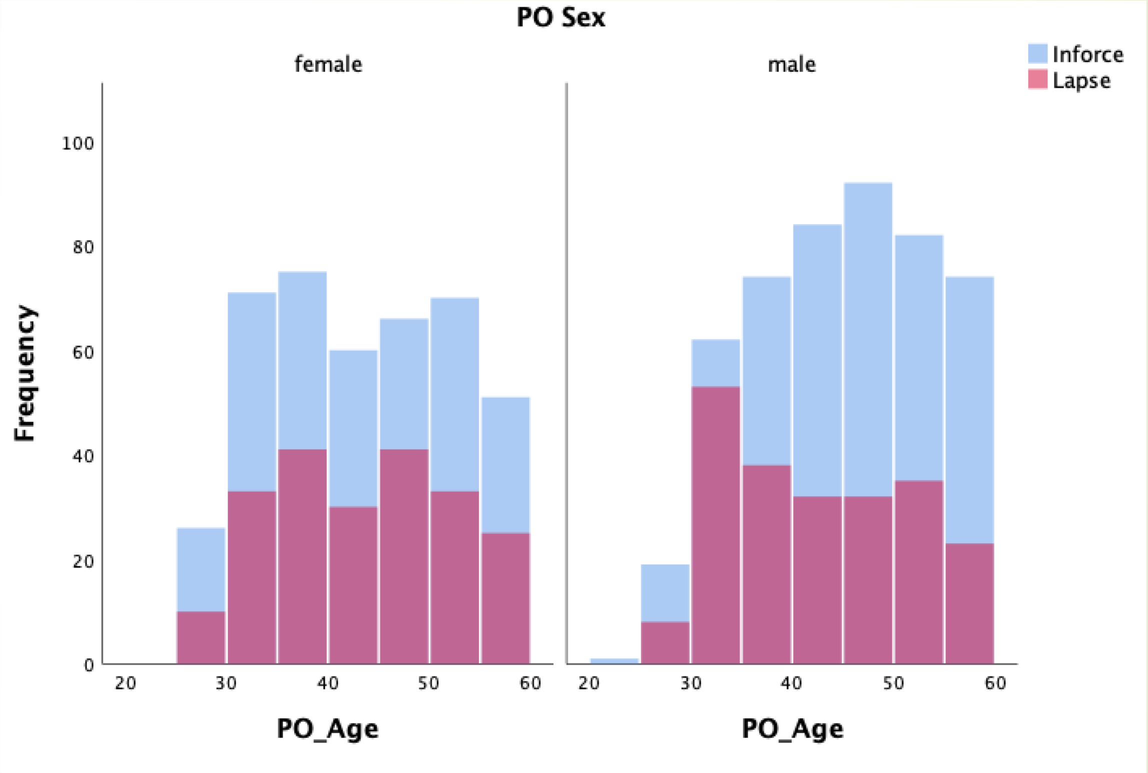




- Others Distribution channels and Bancas have the highest and the lowest lapsed rate in Distribution channel.
- Monthly payment mode likely have the highest lapse rate among 4 types of payments.
- Seem that year of experience of agent has correlation to lapse rate.
- Policy with high coverage is likely to lapse than others.

Age and Gender

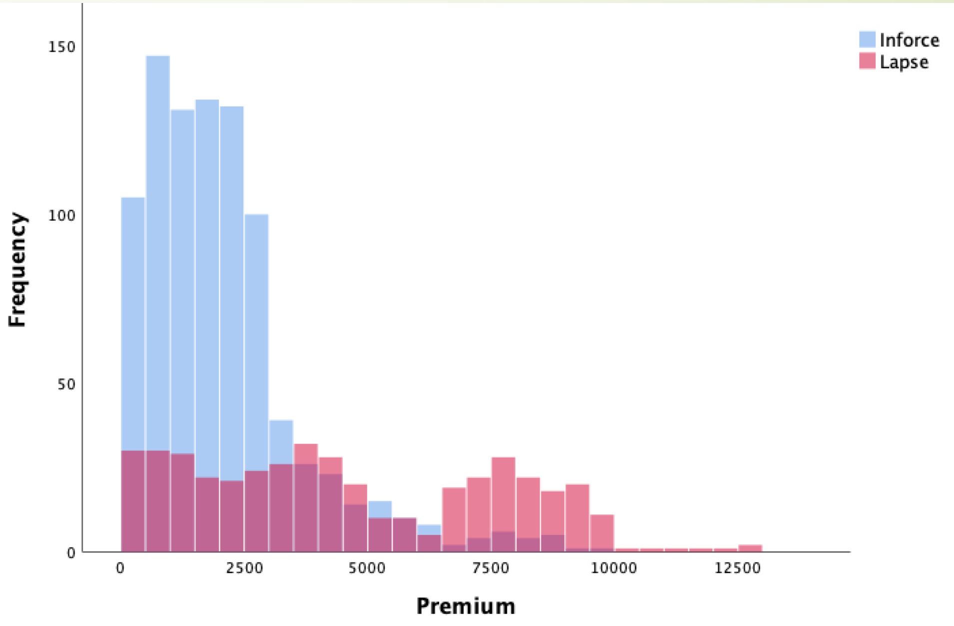
- Lapse rate for male is higher than female for age from 30-35.
- In other age range, lapse rate for female is likely higher than male.

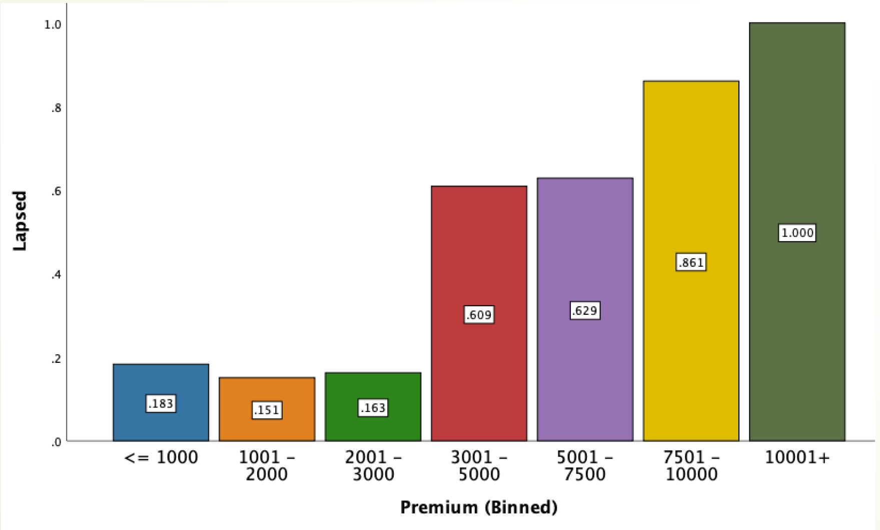


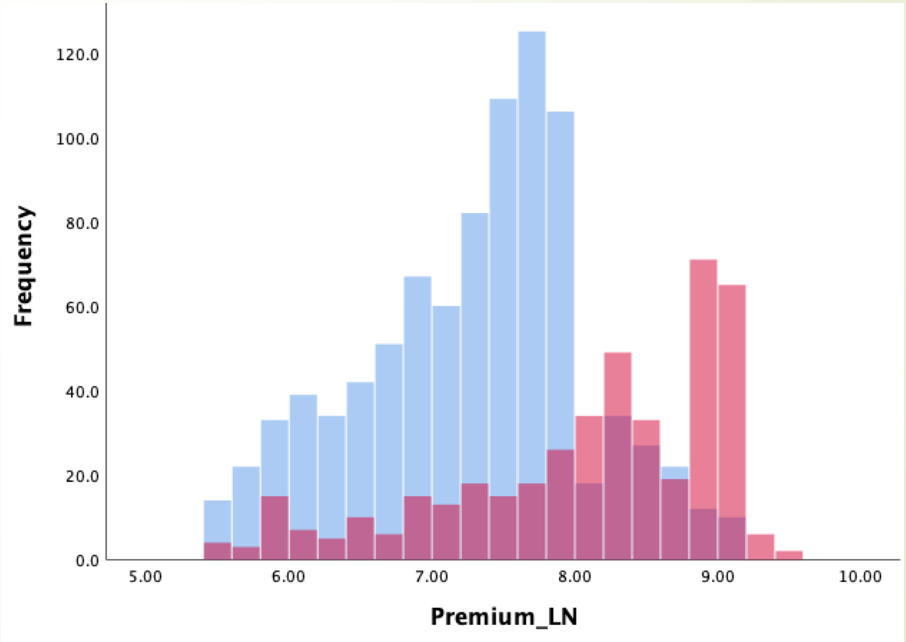
Premium

- Policies with premium less than 3000 are more likely inforce. Inforce rate significantly reduce for policy with premium higher than 3000.
- To handle continuous variable *Premium*, we transform *Premium* into:
  - *Premium (Binned)*
  - *Premium_Ln* = LN(Premium)

  which will be used to build the model as replacing for original *Premium*







Modeling

- We will build the Logistic Regression model with two studies:
  - Study 1 : Analyze Lapse with Customer Demography and Policy Detail
  - Study 2 : Analyze Lapse with Customer Event and Interaction

Logistic Regression

Logistic Regression Analysis is a statistical analysis technique to :

- **model** the probability of an event occurring depending on the values of the independent variables
- **estimate** the probability that an event occurs for a randomly selected observation versus the the probability that the event does not occur
- **predict** the effect of a series of variables on a binary response variable
- **classify** observations by estimating the probability that an observation is in a particular category (such as Lapse or No-Lapse in our study).

Study 1:

Analyze Lapse with Customer Demography and Policy Detail

Interpreting the output

- In this study # 1, we analyze the probability of lapse based on the following features: Sex, Age, Occupation, Premium (Binned), Payment Term, Coverage Period, Distribution Channel.
- ***Interpreting the output:***
  - ***Classification table:*** compares the actual and predicted groups to assess how many would be correctly classified.

Block 0: Beginning Block

Classification Table[a,b]

Observed			Predicted: Lapsed: Inforce	Predicted: Lapsed: Lapse	Percentage Correct
Step 0	Lapsed	Inforce	907	0	100.0
		Lapse	434	0	.0
	Overall Percentage				67.6

a. Constant is included in the model.

b. The cut value is .500

Block 1: Method = Enter

Classification Table[a]

Observed			Predicted: Lapsed: Inforce	Predicted: Lapsed: Lapse	Percentage Correct
Step 1	Lapsed	Inforce	829	78	91.4
		Lapse	174	260	59.9
	Overall Percentage				81.2

a. The cut value is .500

The correct classification percentage is now improved after using the fitting the model.

Interpreting the output

- ***Omnibus Tests of Model Coefficients:*** is used to check that the new model (with explanatory variables included) is an improvement over the baseline model
  - $\chi 2$= [-2LL (baseline)] - [-2LL (new)]

- ***Model Summary:***
  - Deviance -2LL: is used to explore how well a logistic regression model fits the data.
  - The R2 values tell us approximately how much variation in the outcome is explained by the model.

Omnibus Tests of Model Coefficients

		Chi-square	df	Sig.
Step 1	Step	453.880	16	.000
	Block	453.880	16	.000
	Model	453.880	16	.000

The model was statistically significant when compared to the null model, $\chi 2(16) = 453.880$, $p < 0.001$.

Model Summary

Step	-2 Log likelihood	Cox & Snell R Square	Nagelkerke R Square
1	1234.660[a]	.287	.401

a. Estimation terminated at iteration number 20 because maximum iterations has been reached. Final solution cannot be found.

The Nagelkerke R Square value is 0.401 so 40.1% of the variation in outcome can be explained by the full model suggesting that predictions are fairly reliable.
Between 29% and 40% of the variance of dependent variable is explained by our independent variables (aka our model)

Interpreting the output

Hosmer and Lemeshow Test

Step	Chi-square	df	Sig.
1	14.990	8	.059

Contingency Table for Hosmer and Lemeshow Test

		Lapsed = Inforce		Lapsed = Lapse		
		Observed	Expected	Observed	Expected	Total
Step 1	1	120	126.675	14	7.325	134
	2	122	121.250	12	12.750	134
	3	116	117.195	18	16.805	134
	4	116	113.366	18	20.634	134
	5	103	108.860	31	25.140	134
	6	107	103.489	27	30.511	134
	7	102	93.465	32	40.535	134
	8	73	66.197	61	67.803	134
	9	35	39.266	99	94.734	134
	10	13	17.237	122	117.763	135

*The **Hosmer and Lemeshow Test** of the goodness of fit suggests the model is a good fit to the data as p=0.059 – insignificant value*

- ***Hosmer and Lemeshow Test***
  - Is a test for *Goodness of fit* for logistic regression model. A goodness of fit test tells you how well your data fits the model. Specifically, the HL test calculates if the observed event rates match the expected event rates in population subgroups.
  - The output returns a **chi-square** value (a Hosmer-Lemeshow chi-squared) and a **p-value** (e.g. Pr > ChiSq). *Small p-values mean that the model is a poor fit*.
  - In HL test we want p>.05 , insignificant values.
    - The higher p-value, the good fit of model.

Interpreting the output

- **Variables in the Equation**
  - The **Exp(B)** - Odds is the Ratio of Probability **P(A)/P(B)**

    P(A) Probability of falling into target group; P(B): Probability of falling into the non-target group

    - Exp(B)=1: then P(A)=P(B) - No relationship between predictor (or IV) and response (DV)
  - Exp(B) >1: ***(Probability of Event Occurring)*** : then P(A)>P(B),Event is likely to occur. Essentially a positive relationship (positive coefficient B>0)
  - Exp(B) <1 : ***(Probability of Event Occurring Decrease)*** : then P(A)<P(B),Event is unlikely to occur. then P(A)<P(B) : A negative regression coefficient (B<0)

The logistic equation is :

- log(p/1-p) = b0 + b1*x1 + b2*x2 + b3*x3 + b3*x3+b4*x4

where p is the probability of being lapsed.

- log(1/p) = 21.405+1.378*Occupation(1)+1.426*Occupation(2)+ 0.865*Occupation(3)-0.829*CoveragePeriod(1)-0.411*CoveragePeriod(2)+0.603*PaymentTerm(1)…

Variables in the Equation

		B	S.E.	Wald	df	Sig.	Exp(B)	95% C.I.for EXP(B)	
								Lower	Upper
Step 1[a]	Occupation			29.550	3	.000			
	Occupation(1)	1.373	.321	18.316	1	.000	3.947	2.105	7.403
	Occupation(2)	1.426	.298	22.957	1	.000	4.162	2.323	7.458
	Occupation(3)	.865	.303	8.128	1	.004	2.376	1.311	4.306
	CoveragePeriod			16.818	2	.000			
	CoveragePeriod(1)	-.829	.206	16.204	1	.000	.437	.292	.654
	CoveragePeriod(2)	-.411	.161	6.494	1	.011	.663	.483	.909
	PaymentTerm			9.217	3	.027			
	PaymentTerm(1)	.603	.215	7.842	1	.005	1.828	1.199	2.789
	PaymentTerm(2)	.424	.182	5.421	1	.020	1.528	1.069	2.183
	PaymentTerm(3)	.298	.210	2.016	1	.156	1.348	.893	2.034
	INS_Age	-.017	.007	6.465	1	.011	.983	.970	.996
	Premium (Binned)			223.454	6	.000			
	Premium (Binned)(1)	-22.672	14748.348	.000	1	.999	.000	.000	.
	Premium (Binned)(2)	-22.775	14748.348	.000	1	.999	.000	.000	.
	Premium (Binned)(3)	-22.557	14748.348	.000	1	.999	.000	.000	.
	Premium (Binned)(4)	-20.750	14748.348	.000	1	.999	.000	.000	.
	Premium (Binned)(5)	-20.714	14748.348	.000	1	.999	.000	.000	.
	Premium (Binned)(6)	-19.393	14748.348	.000	1	.999	.000	.000	.
	PO_Age	-.016	.008	3.746	1	.053	.984	.969	1.000
	Constant	21.405	14748.348	.000	1	.999	1.977E+9		

a. Variable(s) entered on step 1: Occupation, CoveragePeriod, PaymentTerm, INS_Age, Premium (Binned), PO_Age.

- *Occupation*: Policy with occupation class 1 & 2 are about 3.95 and 4.16 times more likely to lapse than those in class 4 (reference class).
- *Coverage*: Policy with 1-5 years and 5-10 years Coverage Period have the negative coefficient so they were less likely to lapse as 56% and 34% as comparing to 10-20+ years coverage.
- *INS_Age* (p=.011) has a negative coefficient so policy which have higher INS_Age were less likely to lapse.
- *PO_Age* (p=.053) did not add significantly to the model
- *Premium (Binned):* Only group with premium higher than 10.000 were statistically significant to the model.

Interpreting the output

Block 1: Method = Enter

Classification Table[a]

Observed			Predicted: Lapsed Inforce	Predicted: Lapsed Lapse	Predicted: Percentage Correct
Step 1	Lapsed	Inforce	829	78	91.4
		Lapse	174	260	59.9
	Overall Percentage				81.2

a. The cut value is .500

- Using our model, it is accurated predict 81.2%
  - ***Specificity*** or True Negative Rate is 91.4%
  - ***Sensitivity*** or True Possitive Rate is 59.9%
- PAC (percentage accuracy in classification)=81.2%

Reporting Logistic Regression

- *The overall model was statistically significant when compared to the null model, $\chi2(16, N=1341) = 453.880$, $p < 0.001$, explained 40.1% of the variation of dependent variable (Nagelkerke R2) and correctly classified 81.2% of cases. Occupation, Coverage Period, Payment Term , INS_Age and Premium group (with premium above10,000) were significantly predict the model but Sex (both PO and INS), PO_Age, and other Premium groups (less than 10,000) were not.*

Study 2:

Analyze Lapse with Customer Event and Customer Interaction

The model

- Maintaining customer satisfaction by interacting with customer at their events is crucial to keep customer loyalty.
- In this study, logistic regression is used to analyze the relationship between predictors: *NumOfReinstated, NumOfClaims, NumOfEmails, NumOfCalls, DistributionChannel, AgentYearSVR, Premium_LN* and reponse varaible *Lapsed*.
- **Classification Table** for logistic regression model:

Classification Table[a]

Observed			Predicted Lapse: Inforce	Predicted Lapse: Lapse	Percentage Correct
Step 1	Lapse	Inforce	821	81	91.0
		Lapse	187	245	56.7
	Overall Percentage				79.9

a. The cut value is .500

- This model correctly predict 79.9%
  - ***Specificity*** or True Negative Rate is 91.0%
  - ***Sensitivity*** or True Positive Rate is 56.7%

(*) Note that: with business objective is to improve Inforce rate, the higher Specificity the better.

The model

Omnibus Tests of Model Coefficients

		Chi-square	df	Sig.
Step 1	Step	427.614	26	.000
	Block	427.614	26	.000
	Model	427.614	26	.000

Model Summary

Step	-2 Log likelihood	Cox & Snell R Square	Nagelkerke R Square
1	1252.502[a]	.274	.383

a. Estimation terminated at iteration number 5 because parameter estimates changed by less than .001.

Hosmer and Lemeshow Test

Step	Chi-square	df	Sig.
1	26.402	7	.000

- **Omibus Test of Model Coefficients**
  - *$\chi 2(20, N=1341) = 427.61$, $p < 0.001$, we have a significant model.*
- **Model Summary**
  - *From 27.4% to 38.3% of the variance in the dependent variable is explained by the model.*
- **Hosmer and Lemeshow**
  - *$p<0.05$, it's significant value. The model does not fit the data.*

The model

- Using ***Variables in the Equation*** to fomulate the logistic equation:
  - **Y = log(1/p)** = -13.14 + 0.39*NumOfReinstated(1) + 1.33*NumOfReinstated(2) + 2.44*NumOfReinstated(3) + … + 0.17*NumOfClaims(1) +… - 0.26*NumOfEmails(1) +…. – 0.49*NumOfCalls(1) - … - 1.52*NumOfCalls(5) - … -0.23*AgentYearSVR(2) - … - 2.12*AgentYearSVR(5) +….+1.36*Premium_LN

And then, we can calculate **p** = $\frac{e^Y}{1+e^Y}$

where p is the probability of being lapsed.

Variables in the Equation

		B	S.E.	Wald	df	Sig.	Exp(B)	95% C.I.for EXP(B) Lower	Upper
Step 1[a]	NumOfReinstated			93.552	5	.000			
	NumOfReinstated(1)	.396	.198	4.018	1	.045	1.486	1.009	2.189
	NumOfReinstated(2)	1.334	.238	31.293	1	.000	3.796	2.379	6.058
	NumOfReinstated(3)	2.447	.308	63.274	1	.000	11.555	6.323	21.118
	NumOfReinstated(4)	2.454	.532	21.274	1	.000	11.631	4.100	32.993
	NumOfReinstated(5)	3.599	1.153	9.746	1	.002	36.553	3.817	350.061
	NumOfClaims			41.966	5	.000			
	NumOfClaims(1)	.176	.237	.550	1	.458	1.193	.749	1.899
	NumOfClaims(2)	.814	.241	11.384	1	.001	2.257	1.406	3.620
	NumOfClaims(3)	1.728	.372	21.563	1	.000	5.629	2.714	11.673
	NumOfClaims(4)	2.204	.492	20.081	1	.000	9.065	3.456	23.773
	NumOfClaims(5)	.484	.803	.364	1	.546	1.623	.337	7.828
	NumOfEmails			18.774	5	.002			
	NumOfEmails(1)	-.267	.192	1.935	1	.164	.766	.526	1.115
	NumOfEmails(2)	-.194	.281	.477	1	.490	.823	.474	1.429
	NumOfEmails(3)	-1.193	.292	16.666	1	.000	.303	.171	.538
	NumOfEmails(4)	-.948	.615	2.379	1	.123	.387	.116	1.292
	NumOfEmails(5)	-.466	.578	.652	1	.419	.627	.202	1.946
	NumOfCalls			37.265	5	.000			
	NumOfCalls(1)	-.493	.165	8.893	1	.003	.611	.442	.845
	NumOfCalls(2)	-1.096	.266	17.005	1	.000	.334	.199	.563
	NumOfCalls(3)	-1.511	.294	26.451	1	.000	.221	.124	.392
	NumOfCalls(4)	-1.618	.585	7.655	1	.006	.198	.063	.624
	NumOfCalls(5)	-1.529	.599	6.522	1	.011	.217	.067	.701
	AgentYearSVR			20.006	5	.001			
	AgentYearSVR(1)	.134	.167	.641	1	.423	1.143	.824	1.586
	AgentYearSVR(2)	-.230	.196	1.373	1	.241	.795	.541	1.167
	AgentYearSVR(3)	-.994	.420	5.599	1	.018	.370	.162	.843
	AgentYearSVR(4)	-.596	.458	1.693	1	.193	.551	.225	1.352
	AgentYearSVR(5)	-2.122	.693	9.366	1	.002	.120	.031	.466
	Premium_LN	1.364	.095	204.465	1	.000	3.913	3.246	4.718
	Constant	-11.025	.778	200.581	1	.000	.000		

a. Variable(s) entered on step 1: NumOfReinstated, NumOfClaims, NumOfEmails, NumOfCalls, AgentYearSVR, Premium_LN.

Report & Prediction

Report

Study 1: **Analyze Lapse with Customer Demography and Policy Detail**

- Logistic regression was performed to ascertain the effects of *Sex, Age, Occupation, Premium (Binned), Payment Term, Coverage Period, Distribution Channel* on the likelihood that policy status change *(Inforce versus Lapse)*. The logistic regression model was statistically significant, $\chi2(16, N=1341) = 453.880, p < 0.001$. The model explained *40.1% (Nagelkerke R2)* of the variance in policy status and correctly classified *81.2%* of cases. It was found that:
  - Holding another variable constant, the odds of lapse increase by 294% (95% CI[1.10,6.40]), 316%(95% CI[1.32,6.45]) and 137% (95% CI[.31,3.30]) for policy with occupation class 1,2 and 3 compared to policy having occupation class 4.
  - Holding another variable constant, the odds of lapse decrease 56% (95% CI[.35,.61]) and 34% (95% CI[.10,.52]) if coverage period change from 1-5 years, 5-10 years to 10-20+years.
  - Holding another variable constant, the odds of lapse increase 82% (95% CI[.20,.79] and 52% (95% CI[.07,1.18] for policy with Monthly, Quarterly payment compared to Annually payment.
  - Holding another variable constant, the odds ratio of lapse decrease 2% (95% CI[.01,.03] for each additional INS_Age
  - Sex (both PO and INS), PO_Age, Distribution Channel and other Premium groups (less than 10,000) did not add significantly to the model.

Variables in the Equation

		B	S.E.	Wald	df	Sig.	Exp(B)	95% C.I.for EXP(B) Lower	Upper
Step 1[a]	Occupation			29.550	3	.000			
	Occupation(1)	1.373	.321	18.316	1	.000	3.947	2.105	7.403
	Occupation(2)	1.426	.298	22.957	1	.000	4.162	2.323	7.458
	Occupation(3)	.865	.303	8.128	1	.004	2.376	1.311	4.306
	CoveragePeriod			16.818	2	.000			
	CoveragePeriod(1)	-.829	.206	16.204	1	.000	.437	.292	.654
	CoveragePeriod(2)	-.411	.161	6.494	1	.011	.663	.483	.909
	PaymentTerm			9.217	3	.027			
	PaymentTerm(1)	.603	.215	7.842	1	.005	1.828	1.199	2.789
	PaymentTerm(2)	.424	.182	5.421	1	.020	1.528	1.069	2.183
	PaymentTerm(3)	.298	.210	2.016	1	.156	1.348	.893	2.034
	INS_Age	-.017	.007	6.465	1	.011	.983	.970	.996
	Premium (Binned)			223.454	6	.000			
	Premium (Binned)(1)	-22.672	14748.348	.000	1	.999	.000	.000	.
	Premium (Binned)(2)	-22.775	14748.348	.000	1	.999	.000	.000	.
	Premium (Binned)(3)	-22.557	14748.348	.000	1	.999	.000	.000	.
	Premium (Binned)(4)	-20.750	14748.348	.000	1	.999	.000	.000	.
	Premium (Binned)(5)	-20.714	14748.348	.000	1	.999	.000	.000	.
	Premium (Binned)(6)	-19.393	14748.348	.000	1	.999	.000	.000	.
	PO_Age	-.016	.008	3.746	1	.053	.984	.969	1.000
	Constant	21.405	14748.348	.000	1	.999	1.977E+9		

a. Variable(s) entered on step 1: Occupation, CoveragePeriod, PaymentTerm, INS_Age, Premium (Binned), PO_Age.

Report

Variables in the Equation

		B	S.E.	Wald	df	Sig.	Exp(B)	95% C.I.for EXP(B) Lower	Upper
Step 1[a]	NumOfReinstated			93.552	5	.000			
	NumOfReinstated(1)	.396	.198	4.018	1	.045	1.486	1.009	2.189
	NumOfReinstated(2)	1.334	.238	31.293	1	.000	3.796	2.379	6.058
	NumOfReinstated(3)	2.447	.308	63.274	1	.000	11.555	6.323	21.118
	NumOfReinstated(4)	2.454	.532	21.274	1	.000	11.631	4.100	32.993
	NumOfReinstated(5)	3.599	1.153	9.746	1	.002	36.553	3.817	350.061
	NumOfClaims			41.966	5	.000			
	NumOfClaims(1)	.176	.237	.550	1	.458	1.193	.749	1.899
	NumOfClaims(2)	.814	.241	11.384	1	.001	2.257	1.406	3.620
	NumOfClaims(3)	1.728	.372	21.563	1	.000	5.629	2.714	11.673
	NumOfClaims(4)	2.204	.492	20.081	1	.000	9.065	3.456	23.773
	NumOfClaims(5)	.484	.803	.364	1	.546	1.623	.337	7.828
	NumOfEmails			18.774	5	.002			
	NumOfEmails(1)	-.267	.192	1.935	1	.164	.766	.526	1.115
	NumOfEmails(2)	-.194	.281	.477	1	.490	.823	.474	1.429
	NumOfEmails(3)	-1.193	.292	16.666	1	.000	.303	.171	.538
	NumOfEmails(4)	-.948	.615	2.379	1	.123	.387	.116	1.292
	NumOfEmails(5)	-.466	.578	.652	1	.419	.627	.202	1.946
	NumOfCalls			37.265	5	.000			
	NumOfCalls(1)	-.493	.165	8.893	1	.003	.611	.442	.845
	NumOfCalls(2)	-1.096	.266	17.005	1	.000	.334	.199	.563
	NumOfCalls(3)	-1.511	.294	26.451	1	.000	.221	.124	.392
	NumOfCalls(4)	-1.618	.585	7.655	1	.006	.198	.063	.624
	NumOfCalls(5)	-1.529	.599	6.522	1	.011	.217	.067	.701
	AgentYearSVR			20.006	5	.001			
	AgentYearSVR(1)	.134	.167	.641	1	.423	1.143	.824	1.586
	AgentYearSVR(2)	-.230	.196	1.373	1	.241	.795	.541	1.167
	AgentYearSVR(3)	-.994	.420	5.599	1	.018	.370	.162	.843
	AgentYearSVR(4)	-.596	.458	1.693	1	.193	.551	.225	1.352
	AgentYearSVR(5)	-2.122	.693	9.366	1	.002	.120	.031	.466
	Premium_LN	1.364	.095	204.465	1	.000	3.913	3.246	4.718
	Constant	-11.025	.778	200.581	1	.000	.000		

a. Variable(s) entered on step 1: NumOfReinstated, NumOfClaims, NumOfEmails, NumOfCalls, AgentYearSVR, Premium_LN.

<u>*Study 2:*</u> **Analyze Lapse with Customer Event and Customer Interaction**

- Logistic regression was used to analyze the relationship between predictors *NumOfResinstated, NumOfClaims, NumOfEmails, NumOfCalls, Distribution Channel, AgentYearSVR, Premium_LN* and response variable *Policy Status (InForce vs Lapse)*. The logistic regression model was statistically significant, *$\chi 2(20, N=1341) = 427.61$, $p < 0.001$*. The model explained *38.8% (Nagelkerke R2)* of the variance in the response variable and correctly classified *79.9%* of cases. It was found that:
  - When Exp(B) is greater than 1, increasing values of the variable correspond to increasing odds of *lapse* (the event's occurrence).
    - *NumOfReinstated, NumOfClaims, AgentYearSVR* have the increased odds of lapse compared to their reference category. (except for category with Sig. > 0.05 which is not useful to the model)
    - Premium_LN : increasing premium_LN (~2.71 premium) correspond with increase odds of lapse.
  - When Exp(B) is less than 1, increasing values of the variable correspond to decreasing odds of *lapse* (the event's occurrence).
    - NumOfEmails, NumOfCalls: have the decreased odds of lapse compared to their category.

Prediction

- In Prediction, we will interest in studying the case which are still inforce but being predicted as Lapse. They are the **False Positive ('FP')** cases in Classification table.
- There are **78 cases** by Model 1 ('M1') and **81 cases** by Model 2 ('M2'). By joining 2 list, we have **38 FP cases** predicting by 2 models which we should pay more attention to prevent futher lapse.

False Positive Cases

	Policy Num	Policy Status	Predicted group	Occupation	CoveragePeriod	PaymentTerm	Premium	NumOf Reinstated	NumOf Claims	NumOf Calls	NumOf Emails	Agent YearS VR	Premium_ LN	Predicted probability by M1	Predicted probability by M2
1	637	Inforce	Lapse	Class 1	10-20+ years	Annualy	7540	2	0	1	1	3	8.93	.90855	.81777
2	77	Inforce	Lapse	Class 3	10-20+ years	Quartely	7890	0	0	1	1	2	8.97	.90703	.64399
3	644	Inforce	Lapse	Class 3	5-10 years	Quartely	8900	1	0	5	2	2	9.09	.87739	.54697
4	101	Inforce	Lapse	Class 3	10-20+ years	Quartely	8230	2	1	2	1	2	9.02	.87540	.82594
5	34	Inforce	Lapse	Class 1	10-20+ years	Annualy	8750	0	0	1	0	3	9.08	.84663	.65406
6	3	Inforce	Lapse	Class 3	10-20+ years	Monthly	8890	0	0	1	1	2	9.09	.83972	.68038
7	833	Inforce	Lapse	Class 3	5-10 years	Quartely	7900	0	0	0	1	2	8.97	.82779	.74784
8	1020	Inforce	Lapse	Class 2	10-20+ years	Semi-annual	5204	0	0	0	1	3	8.56	.81461	.53846
9	651	Inforce	Lapse	Class 3	5-10 years	Quartely	9600	0	0	1	1	3	9.17	.79762	.62174
10	839	Inforce	Lapse	Class 3	5-10 years	Annualy	8280	1	0	1	0	1	9.02	.79268	.76624
11	865	Inforce	Lapse	Class 3	5-10 years	Quartely	7560	0	3	3	1	2	8.93	.78951	.77620
12	43	Inforce	Lapse	Class 3	1-5 years	Quartely	9250	0	0	1	0	1	9.13	.75991	.71957
13	1007	Inforce	Lapse	Class 4	10-20+ years	Quartely	8800	0	0	0	3	1	9.08	.75026	.54350
14	354	Inforce	Lapse	Class 2	5-10 years	Monthly	4810	3	0	3	3	3	8.48	.74215	.51411
15	1196	Inforce	Lapse	Class 2	5-10 years	Monthly	3768	3	0	1	1	1	8.23	.73950	.86963
16	694	Inforce	Lapse	Class 1	10-20+ years	Annualy	6471	0	0	0	1	2	8.78	.71217	.69315
17	664	Inforce	Lapse	Class 3	10-20+ years	Quartely	4227	3	1	1	2	2	8.35	.70522	.91958
18	290	Inforce	Lapse	Class 2	5-10 years	Monthly	5199	1	0	1	1	1	8.56	.69718	.57096
19	859	Inforce	Lapse	Class 2	5-10 years	Quartely	3644	4	0	1	4	3	8.20	.69529	.72059
20	807	Inforce	Lapse	Class 1	5-10 years	Monthly	5695	3	0	3	1	1	8.65	.68777	.80884

Top 20 FP cases with higher lapse predicted probability

Summary of Findings and Suggestions

- Client in Occupation Class 1 are more likely to Lapse than class 2,3 and 4, we should take more good care to customer in occupation class 1
- Policy with 1-5 years and 5-10 years Coverage Period were less likely to lapse as 56% and 34% as comparing to 10-20+ years coverage, should we take further study with additional variable Policy Year.

- *As prediction, we can filter FP cases which their lapse possibility rate higher than 75%. Although, we can not change customer demographics, policy detail but we can improve customer interaction, change servicing agent in order to reduce the possibility rate.*
- *Either we apply Model 1 or Model 2 to predict the probability of Lapse for future policy data, they can reach 79.9% to 81.2% accuracy.*

Thank you