Capstone Project
Background and Objectives

LIFE INSURANCE:POLICY LAPSE MODEL

DATA SCIENCE
INSTITUTE

Background

1. Background

Life insurance persistency data refers to information about the continuity of policyholders' premiums payments over time. By analyzing this data using data science techniques such as machine learning algorithms and predictive modeling, we can identify patterns and factors associated with lapses in policy payments. By leveraging features such as policyholder demographics, payment history, policy attributes, and external factors, we can develop models to predict the likelihood of policy lapses. These models can help insurers proactively identify at-risk policies and take preventive actions to mitigate lapses, thereby improving overall persistency rates.

TA SCIENCE
INSTITUTE

Background

2. Primary Objectives:

1. To study association between policy lapse status and various factors such as demographics and policy details
2. To develop predictive model for estimating probability of policy lapse
3. Develop a dashboard to accept values of input variables and predict the probability of policy lapse

Background

3. Data

The following datasets are available:
1. Customer Info
2. Policy Details
3. Policy Status

Data : Customer Info

Content

The dataset includes customer information such as unique identifiers, age, gender, rewards and so on

PolicyID	Loyalty	Gender	Source	Age	N_policies	N_years	Marital_status	NPS	payment_freq	Familysize
PID - 001	No	M	Digital	38	3	1	Single	3	Monthly	3
PID - 002	No	M	Agent	55	4	6	Single	7	Quarterly	2
PID - 003	No	F	Digital	33	5	3	Married	1	Monthly	5
PID - 004	No	M	Agent	20	1	4	Single	8	Yearly	2
PID - 005	No	M	Agent	39	1	7	Married	8	Monthly	1
PID - 006	No	M	Agent	30	3	2	Married	5	Yearly	2

Columns	Description	Type	Possible values
PolicyID	Policy Id	Alpha numeric	-
Loyalty	Member of Loyalty Program	Factor	Yes or No
Gender	Gender	Factor	M or F
Source	Source	Factor	Digital or Agent
Age	Age	Numeric	-
N_policies	Number of policies	Numeric	-
N_years	Number of Years with Company	Numeric	-
Marital_status	Marital Status	Factor	Marries/Single
NPS	Net Promoter Score	Numeric	1-10
Payment_freq	Payment Frequency	Factor	Monthly, quarterly, half-yearly, yearly
Familysize	Family size	Numeric	-

Data : Policy Details

Content

This data presents customer policy details such as policy id, sum insured and issue date.

PolicyID	Sum_insured	Issue_date
PID - 001	8269994	21-02-2023
PID - 002	2225474	24-01-2023
PID - 003	6740199	19-02-2023
PID - 004	4407422	02-02-2023
PID - 005	7413919	24-01-2023
PID - 006	3918066	21-01-2023
PID - 007	4141953	06-01-2023
PID - 008	4019816	24-01-2023
PID - 009	4316581	19-01-2023

Columns	Description	Type	Possible values
PolicyID	Policy Id	Alpha numeric	-
Sum_insured	Sum insured	Numeric	-
Issue_date	Policy issue date	Date	

Data : Policy Status

Content

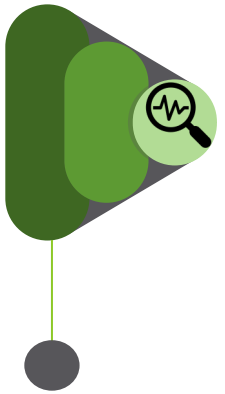
The dataset includes Policy status for each policy holders, lapsed or not

PolicyID	Policy_Status
PID - 001	1
PID - 002	0
PID - 003	1
PID - 004	0
PID - 005	0
PID - 006	1
PID - 007	1

Columns	Description	Type	Possible values
PolicyID	Policy Id	Alpha Numeirc	
Policy_Status	Policy status	Factor	1:Lapsed, 0: Not Lapsed

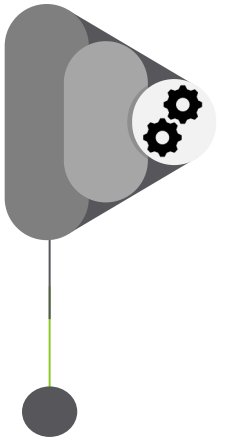
Next steps

Data management



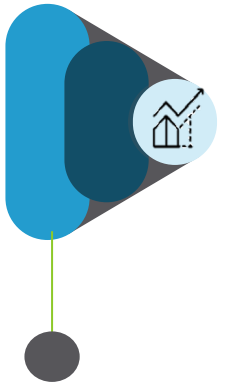
- Compile all 3 data files based on Policy ID
- Data cleaning , Handling missing values and completing Basic Data checks
- Check if any variables needed to be feature coded i.e made into groups or want to be left as continuous variables

Descriptive Statistics & Data visualization



- Understanding the data better like lapse rate.
- How can this data be presented better visually ?
- Once again post Data visualization check if any variable needs to be feature coded

Predictive modelling



- Answer the objectives
  a) To identify patterns and factors associated with lapses in policy payments.
- Using different Predictive model techniques to find Significant variables
- Ensure you follow all steps like Train and test data , checking for Multicollinearity
- Check if any other ML technique fits better