**Health Insurance Dataset Analysis CRISP-DM**

**1. Business Understanding:**

**Business Goal/Objective:**  
The main aim is to identify patterns and trends that may inform business strategies around product design, distribution, risk selection and customer retention.

**Key Business Questions:**

1. Product Design - Which policy types or product characteristics (individual vs. collective, dental vs. standard vs. premium, etc.) show higher claim ratios, Persistency rates, lapse (fallout) rates and what customer segments are driving these patterns?
2. Distribution - How do lapse rates, claim costs, and customer profiles vary across distribution channels (agency, direct, intermediary) and which channel delivers the most profitable and persistent business?
3. What demographic (age, gender), socioeconomic (income, education), and geographic (municipality, province) factors are associated with higher risk of claims or early lapse? How does this affect our policy and claim underwriting procedures?
4. What characteristics differentiate policies that lapse early (before expiration) from those that remain active till the policy term ends and what patterns suggest opportunities for intervention to improve persistency?

**2. Data Understanding.**

The dataset comes from a secondary source, Mendeley dataset covering 3 years (2017–2019), **228,711 rows × 42 columns**.

Each row represents an insured (individual) policy, while each column represents a distinct variable, demographics, policy type, premiums, claims, geography, etc.

The following table is a summary of all the columns and what they entail:

|  |  |
| --- | --- |
| **Initial Variables of Interest:** | **Columns** |
| **Dates & Durations** | date\_effect\_insured, date\_lapse\_insured, seniority\_insured |
| **Unique identification** | ID, ID\_Policy, ID\_Insured |
| **Policy Status** | lapse, exposure\_time, period, new business |
| **Marketing** | Distribution\_channel |
| **Insurance Penetration** | n\_insured\_pc , n\_insured\_mun, n\_insured\_prov, IICIMUN, IICIPROV, C\_H, C\_GI, C\_II,C\_IE\_P ,C\_IE\_S , C\_IE\_T, C\_GE\_P ,C\_GE\_S, C\_GE\_T, C\_C |
| **Demographics** | age, gender, C\_GI, C\_II, C\_H |
| **Financials** | premium, cost\_claims\_year, n\_medical\_services |
| **Product Info** | type\_policy, type\_policy\_dg, type\_product |
| **Utilization** | n\_medical\_services, reimbursement |
| **Geographic Distribution** | n\_insured\_mun, IICIMUN, C\_C |

1. **Checking for Missing values:** concentrated in lapse dates, socioeconomic indices, and geographic classifications.
2. **Checking for Duplicates:** none.
3. **Checking for mismatch in Data types:** mix of categorical, numerical, and datetime fields.
4. **Descriptive analysis:**

Summary statistics for the numerical and categorical columns separately.

Group-wise summaries to discover high-risk or high-cost clusters.

Correlation analysis among numerical variables.

**3. Data Preparation:**

**Load and inspect the data**

Check the Imported dataset (columns, datatypes and basic statistics) for completeness and consistency

Identify data types (numerical, categorical, date/time)

Exploring summary/descriptive statistics

**Clean and transform the data**

* Converting categorical codes to descriptive labels (e.g., I → Individual, A → Agency).
* Assess missing values and outliers

Handling missing values by:

Numerical features filled with **median**.

Categorical features filled with **mode**.

* Classify or group the data into new business vs in force, lapsed vs active, scheme vs individual policies.

**Deriving new features**

* + Policy duration (Duration the policy has been active) = difference between effect and lapse date
  + Policy term (scheme-level) = difference between effect and end date
  + Age groups (e.g., 0–18, 19–35, 36–55, 56+)
  + Policy seniority bins (<1yr, 1–5yrs, 5–10yrs, etc.)
  + Flags for lapsed and new business policies (IsLapsed, IsNewBusiness, Policy Category.)
* Handling categorical variables i.e. Gender, type, distribution channel etc for modelling.
* Calculating lapse rates by:
  + Age, gender, seniority, product type, policy type
  + Region, seniority level
* Create Measures:
  + Average premium per policy type
  + Claim ratios = cost\_claims\_year / premium
  + Medical service rate per age band, gender
  + Reimbursement rate by channel and product

**4. Exploratory Data Analysis**

The main aim is to identify patterns and trends that may inform business strategies around product design, distribution, risk selection and customer retention.

This objective translates into the following questions,

1. Which policy types or product characteristics show higher claim ratios?
2. Which channel delivers the most profitable and persistent business?
3. What demographic, socioeconomic, and geographic factors are associated with higher risk of claims?
4. What characteristics differentiate policies that lapse early from those that remain active till the policy term ends?

**a) Product Design**

* **Highest lapse rates:** International Collective and Premium Collective policies.
* **Highest persistency:** Dental Collective policies.
* **Highest claim ratio:** Premium Collective (≈0.85).
* Insight: Collective international/premium plans drive fallouts; Dental coverage retains customers better.

**b) Distribution Channels**

* **Direct business**: highest lapse rates, but also highest profitability (low claim costs).
* **Intermediaries**: highest average premiums but also highest claims, making them least profitable.
* **Agency**: high premiums, but also relatively high claims.

**c) Demographics & Socioeconomic Factors**

* **Age & gender**: younger groups (19–35) show higher lapse rates, especially males.
* **Income**: lapse decreases as insured income level increases (inverse relationship).
* **Education**: higher education correlates with lower lapse.
* **Geography**: higher lapse rates in lower-income municipalities and regions with lower insurance penetration.

**d) Cluster / Risk Segments**

* High-risk clusters identified:
  + Younger, low-income insureds with <1 year seniority.
  + Collective international and premium plan policies.
* Correlation analysis showed:
  + Strong link between **premium** and **claim costs**.
  + Negative correlation between **seniority** and **lapse**.

**5. Evaluation**

The expectation is that the persistency rates across the population is high or positive.

If the persistency rates are negative or extremely low, then the analysis would be considered erroneous unless a proper reason for such rates is discovered.

The assumption for lapse rates is vice versa; extremely high values should be flagged and a possible reason outlined.

The trends should be reasonable (subjectiveness is required for this).

Is the product design or distribution the reason why policies are exiting prematurely? The result would be to find the feature that isn’t favourable., distribution, risk selection and customer retention

How are the demographic features of the population, affecting customer retention. Are we underwriting risky business? We expect that majority of our population come from areas with a high insurance penetration hence are insurance literate. We also expect that most of our population comes from families that are financially stable. Otherwise, an alarm must be raised about our underwriting and sales processes.

**6.Deliverables**

* **Interactive dashboard** (Power BI)
  + Filters by product type, region, channel, lapse status
  + Visual KPIs and time series trends
* **Insight Report**
  + Summary of trends and anomalies with business recommendations
* **Prepared dataset.** 
  + Cleaned dataset ready for machine learning (persistency model in next phase)

**7. Deployment**

The analysis outputs, power BI dashboard and report will be found in both my portfolio and GitHub accounts linked below.