Abstract:

Due to the long-term coverage a majority of insurance plans entail, insurers are extremely sensitive to changes in regional and international financial markets. The insurance market was slowing down globally, and many businesses were under pressure to produce profits. Longer time periods for claim settlement are seen to be the primary cause of this. The 1 lakh entries in this dataset serve as a case study to examine and evaluate the healthcare coverage claims made by the Ideal Company. The second justification is to work on preventing fraud, which is one of the major sources of industry profit leakage. The key is to prioritise.

Use case:

In the health insurance industry for transformation, it includes many stakeholders and market participants. These are given below:

Consumers:

* The people who suffer from the ailments and be a beneficiary under insurance policy for various health services and treatments.

Professionals:

* The skilled decision makers medically. Their skills, experience matters most.

Health care institutions:

* Majority of the decisions are determined by the choices of institution for healthcare service providers.

Pharmaceutical companies:

* The quality of the medicines and health care devices manufactured influence the healthcare services available in any region.

Healthcare policy makers:

* The regulatory measures for the policy effect the cost and quality of healthcare and the respective insurance service for the individual.

Insurance companies:

* These companies prepare the policy according to the customer’s risk profile.
* They provide coverage for policy holder in return of an insurance premium.

Third-party assistance (TPA):

* These are the mediators between insurance providers and the customers.
* They help in providing information regarding underwriting the policy for its smooth processing of claims.

Agents or Banks:

* The insurance service provider either hires the agents/employees to sell various products or collaborates with financial institutions such as banks and third-party providers to cross-sell insurance products to their existing customer base.

Business Problem:

* Given that insurance fraud is one of the major causes of white-collar crime worldwide, a lot of police work is expended in locating and eradicating it. No insurance sector is immune to fraud, including the health insurance market.
* According to experts, fraud costs the global healthcare industry roughly 6% of its annual budget.
* Finding the fraud that has been committed in any of the aforementioned situations can be a time- and effort-consuming task. The process of looking for fraud can be challenging given that healthcare spending can occur suddenly, urgently, and unpredictably.
* Companies must strike a balance between their financial limitations and the realities of the healthcare system.
* The most prevalent form of health insurance fraud is financial, which includes illegal billing and false claims. It is typically related to organisational and health information management issues.
* Healthcare fraud raises premium costs, adds to consumer costs, and reduces coverage. It raises the cost for employers to provide health insurance to their workers, which has an impact on business expenses. Along with financial losses, fraudulent actions expose people to exploitation and needless, risky medical procedures, which can have grave health consequences.
* However, fraud detection must deal with the potential for false positives. Allowing bogus claims to be processed damages the business's profitability and raises premiums.
* However, requiring legitimate claims to go through the fraud detection procedure raises expenses and detracts from consumer satisfaction. Since these restrictions are mutually exclusive, any attempt to reduce one will typically result in an increase in the opposite side effects.

Data:

The dataset provided consists of Ideal Insurance Company's insurance claims in the area of health insurance. Over 33 criteria are taken into account, and there are almost 1 lakh records in all. The parameters contain facts about the policy, the third-party administrator, and demographic information. The data also includes information about the dates on which the policy was issued, the date on which a claim was made, and the status of that particular claim. The data emphasises study of insurance claims, both fraudulent and not fraudulent.

Data Dictionary:

The attributes of the dataset are as follows:

**Tpa**

* Third-party administration (tpa) ID
* It is the unique code given to each third-party company that holds license for insurance
* In the data set 18 unique tpa code exists
* This is object datatype

**Policy\_ref**

* Policy reference number is a unique identity number issued by the organization for every insurance
* This is object data type

**Member\_id**

* It is the unique number assigned to each member those come under the same insurance
* It is object datatype

**Sex**

* It refers to either Male (M) or Female (F) about the gender of the insurance claimer.
* It is object datatype

**Dob**

* Date of birth of the insurance claimer
* It is date type

**Policy\_start\_dt**

* It is the date from which the insurance has been commenced
* It is date type

**Policy\_end\_dt**

* It is the date on which the insurance expires
* It is date type

**Prod\_code**

* It is the product code refers to the domain for which the insurance is being claimed
* They are categorized into 7 types: Basic Policy (A)

Hospital cash plan (B)

Outpatient coverage (C)

Universal health policy (D)

Microinsurance policy (E)

Package policy (F)

Hybrid Policy (G)

Any other product type (0)

* It is object datatype

**Policy\_type**

* Refres to the type of policy cover taken
* It is categorized into 9 types: Individual (A)

Individual floater (B)

Group (C)

Group floater (D)

Declaration (E)

Declaration floater (F)

Declaration with group insurance (G)

Declaration floater with group insurance (H)

Any other cover type (0)

* It is object datatype

**Sum\_insured**

* The maximum amount from the insurance that is available for the policy holder
* It is int datatype

**Claim\_ref**

* It is the reference number for the insurance claimed
* It is object datatype

**Claim\_dt**

* The date on which the insurance can been claimed information given to the policy holder
* It is date datatype

**Hospital\_id**

* It is the number for each hospital to identify uniquely
* It is object datatype

**Hos\_zipcode**

* The zipcode for the area of hospital location
* It is object datatype

**Admit\_dt**

* The date on which the person is admitted in the hospital
* It is date datatype

**Discharge\_dt**

* The date on which the person is discharged from the hospital
* It is date datatype

**Payment\_dt**

* Date on which the amount claimed from insurance has been settled
* It is date datatype

**Claim\_amt**

* Amount claimed by the policy holder
* It is float datatype

**Nursing\_chg**

* The nursing charges incurred during treatment
* It is float datatype

**Surgery\_chg**

* The surgery charges incurred during treatment
* It is float datatype

**Cons\_fee**

* The doctor consultancy charges incurred during treatment
* It is float datatype

**Test\_chg**

* The prescribed medical test charges incurred during treatment
* It is float datatype

**Pharmacy\_cost**

* The medicine charges consumed during treatment
* It is float datatype

**Other\_chg**

* Anyother charges incurred during treatment apart from above category
* It is float datatype

**Pre\_hsop\_exp**

* Amount claimed for pre-hospitalization treatment
* It is float datatype

**Post\_hosp\_exp**

* Amount claimed for post-hospitalization treatment
* It is float datatype

**Other\_chg\_non\_hosp**

* Other charges that are non-hospital
* This includes laundry, paid TV channels, guest F&B, etc
* This is float datatype

**Copayment**

* The portion of the medical expense that a policyholder needs to pay when he receives medical services
* It is float datatype

**Settle\_amt**

* Final amount paid to the insured or to the medical service provider on behalf of insured
* It is float datatype

**Payment\_type**

* It is mode of payment
* It is categorized into 6 types: Cashless settlement (A)

Reimbursemet to claimant (B)

Cash benefit (C )

Per diem basis (D)

Lump-sum basis (E )

Cashless to the insured (F)

**Hosp\_type**

* To know if the hospital is networked or not
* Networked hospitals are the ones that will provide the benefits if cashless transactions
* So represented by Y (yes) and N (no)
* It is object datatype

**Recommendation**

* It is the recommendation given by the system for the claim settlement according to the policy claim details
* Categorized into 3 types: Green (which means safe)

Discuss (can be issued based on thorough discussion)

Investigate (Proper investigation is required)

* It is object datatype

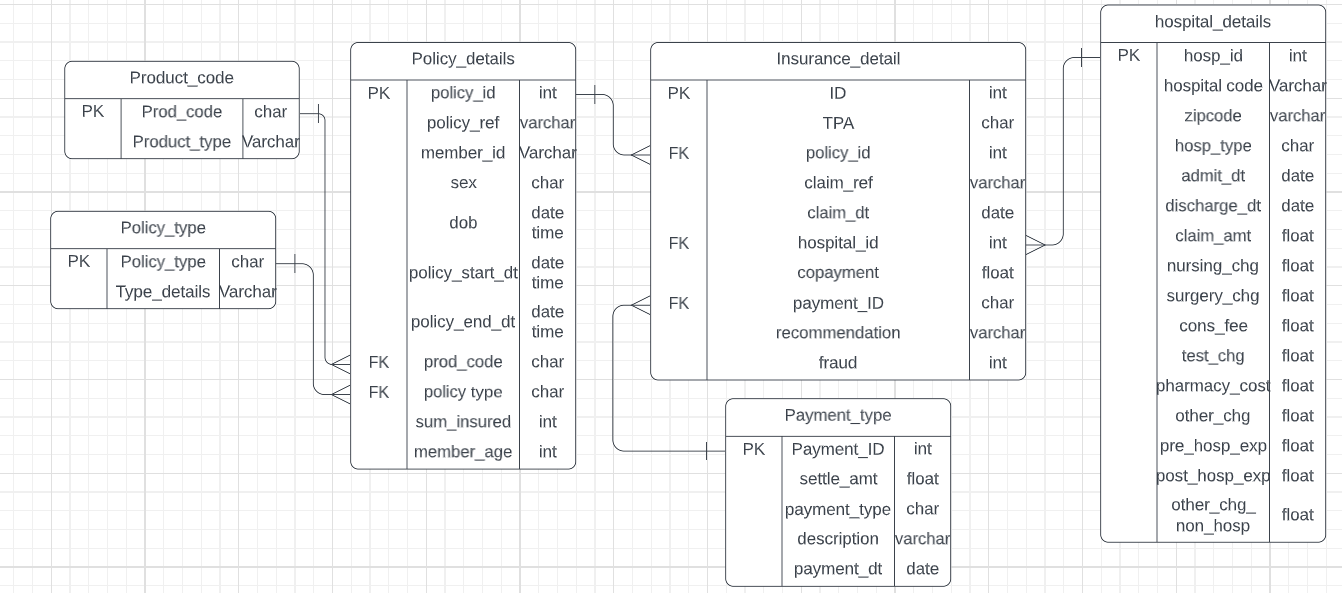
**Fraud**

* The insurance claimed is classified as fraud (1) or non-fraud (0) by the expert consultant
* It is int datatype.

**Questions that can be answered:**

1. Which product code is taken most?
2. What is average difference between the claimed amt and the settled amount
3. What is the average of insured sum?
4. What is the maximum, minimum and average time gap between the claim date and admitted date?
5. How is the claim amount related to the treatment time (admit date and discharge date)?
6. Which policy type is considered the most?
7. Which gender people are diagnosed more M or F?
8. What is the average time taken for the payment after discharge or claim date?
9. Which hospital has more/ less fraud cases?
10. Effect of hospital being networked or not for it be classified as fraud
11. Which charge percentage is more from the claim amount?
12. Is the payment mode is being always cashless when it is a networked hospital?
13. Which type of payment method is used more for the policy?
14. How many policy claims needed Copayment if so what is the claim amount?
15. How many claims are recommended genuine when it is fraud?
16. Which tpa code has more/less fraud cases?
17. What is the average tenure for the policies taken?
18. Relation between sum insured and the policy tenure?
19. Does the policy tenure effect the fraud cases?
20. Multiple claims from a single policy
21. Claim date relation with the policy expiry date

**ER DIAGRAM**



**SQL ASSIGNMENT QUESTIONS TO ANSWER**

The given data set has information regarding insurance claims of ideal insurance company for the health industry. The data is considered to make fraud analysis in comparison with the system recommendation and other parameters.

1. Find out which policy type is considered most for each product code.
2. Get the total insured sum categorised by TPA.
3. Find the statistics of maximum, minimum and average time gap (in days) between the claim dates and admit date.
4. Find the relation between the claim amount and the treatment time (difference of admit date and discharge date).
5. Get the sum of difference between claim amount and settlement amount categorised by payment type.
6. Find the statistics of maximum, minimum and average time gap (in days) between the discharge date and payment date categorised by payment type.
7. What is the average tenure for a policy taken categorised by the product code. (Difference between policy start and end date)
8. Find out which hospital type has claims classified as fraud more either networked or non-networked hospitals.
9. Get the count of policy claims categorised by sex. (Based on member\_id).
10. Get the count of members for each distinct policy reference value.
11. For claims classified as fraud, get the details of claims recommended to be genuine
12. Validate whether the sum of all the charges matches to the claim amount.
13. Find if there are any co-payments made.
14. Which is the most commonly used payment type for the policy claims.
15. What is the average age of the members claiming for the insurance considering all the claim records.
16. Get all the details of claims for the highest sum insured tpa classified as fraud.
17. Add a new column to insurance detail table named as profit which the difference between claim amount and settlement amount.
18. Find out which tpa made highest profit.
19. Find the relationship between sum insured and the policy tenure.
20. For a new policy insurance which tpa, product code and policy type is recommended according to your observation of analysis.

**COMPLETE DATA**

Cleaning up the dataset provided and converting it into sql normalised form ended up the data with the following columns.

Tpa, claim\_ref, caim\_dt, policy\_ref, member\_id, sex, dob,member\_age,policy\_start\_dt,policy\_end\_dt,prod\_code,product\_type,policy\_type\_typedetail,sum\_insured,hospital\_code,hosp\_zipcode,hosp\_type,admit\_dt,discharge\_dt,claim\_amt,nursing\_charge,surgery\_charge,cons\_fee,test\_charge,pharmacy\_cost,other\_charge,pre\_hosp\_charge,post\_hosp\_charge,other\_charge\_non\_hosp,copayment,settle\_amt,payment\_type,payment\_desc,payment\_dt,recommendation,fraud

**TABLEAU ASSIGNMENT**

1. Represent the popular product code and policy type taken
2. Analyse which policy type has more amount of sum insured categorised by sex
3. Average of claim and settlement amount for each tpa
4. Analyse claim amount vs treatment time for each policy type with color for fraud field
5. Analyse claim amount vs treatment time for each policy type with recommendation and fraud fields representation
6. Analyse sum of claim amount vs settle amount for hospital type categorised by fraud
7. Relation for settle amount and hospital type for different payment types
8. Profit analysis for each type categorised by fraud
9. Analysis for recommendation for each tpa categorised by fraud
10. Yearly profit based on difference amount
11. Profit and loss for each tpa