HEALTH INSURANCE CLAIM PREDECTION ANALYSIS

ABSTRACT

This project analyzes health insurance data using Python, SQL, machine learning, and Power BI. It focuses on forecasting claims, optimizing premium pricing, and improving customer retention. Machine learning models predict risks, while SQL facilitates efficient data retrieval. Power BI visualizes insights, enhancing decision-making in the health insurance sector.

Name :- CHIRAG

Contents

1. Introduction and Project Overview

The health insurance project aims to analyze health insurance data to understand trends in coverage, claim amounts, and policyholder behavior. The project utilizes various analytical tools such as Excel, SQL, Power BI, and Python to clean, analyze, and visualize the data. The goal is to create a comprehensive analysis that provides insights into key metrics such as claims, premium distribution, and policy renewal rates.

---

2. Data Sources

The data for this project was collected from a health insurance provider, including information on policyholders, claims, premiums, and coverage details. The dataset includes the following columns:

- **Policyholder ID**
- **Age**
- **Gender**
- **Policy Type**
- **Premium Amount**
- **Claim Amount**
- **Coverage Type**
- **Renewal Status**

The data was obtained in CSV format and stored in a relational database for easy querying and analysis.

---

3. Methodology

- **Data Cleaning:** The data was first cleaned using Python to remove duplicates, handle missing values, and standardize formatting (e.g., ensuring numerical values are formatted correctly). The Python libraries pandas and numpy were used for data manipulation.

- **Data Analysis:** SQL was used to query the relational database and retrieve relevant data. Key insights were derived by analyzing the total premium, claim amounts, and renewal rates. Power BI was used to create dynamic visualizations for better insight delivery.
- **Interpretation:** The data was interpreted by identifying patterns in claims versus premiums, demographic distribution, and policy renewal behavior.

---

4. KPIs and Metrics

- **Total Premium:** Total amount collected for health insurance policies.
- **Claim-to-Premium Ratio:** The ratio of claims paid out to premiums collected.
- **Policy Renewal Rate:** The percentage of policies that are renewed.
- **Average Claim Amount:** The average amount paid out in claims per policyholder.
- **Customer Segmentation:** Breakdown of policyholders based on age, gender, and coverage type.

---

5. Excel Implementation

The health insurance data was analyzed using an Excel workbook, which included the following:

- **Data Import:** Data from CSV was imported into Excel using Power Query.
- **Formulas and Functions:** Key formulas included SUMIF, VLOOKUP, and IF statements to analyze claim amounts, calculate premiums, and determine renewal rates.
- **Pivot Tables:** Pivot tables were created to summarize data, such as total premiums by policy type and average claims by age group.
- **Charts and Visuals:** Charts were created to visualize key metrics, such as bar charts for claims and premium distribution, and line charts to track policy renewal trends.

Python and Machine Learning Implementation

Python Implementation:

In this health insurance project, Python was used for data cleaning, data manipulation, and initial analysis. The key steps involved are:

1. **Data Cleaning:** Python's pandas library was essential for cleaning the data. It allowed us to:
   - **Handle Missing Data:** Any missing or null values in the dataset were identified and dealt with by either removing the rows or imputing the missing values with appropriate methods (e.g., mean, median).
   - **Remove Duplicates:** Duplicate entries were identified and removed to ensure the dataset only contained unique records.
   - **Standardize Data Types:** Python was used to standardize data formats, ensuring that numeric columns (e.g., premium amounts, claim amounts) were properly formatted for analysis.
2. **Data Manipulation and Transformation:** Python allowed for efficient manipulation and transformation of the dataset. Key operations included:
   - **Feature Engineering:** New features were created, such as age groups or premium categories, by transforming existing columns.
   - **Aggregation:** Python was used to aggregate data by different policyholder characteristics, such as age, gender, and policy type, to analyze trends.
3. **Exploratory Data Analysis (EDA):** Python's matplotlib and seaborn libraries were used for initial exploratory data analysis. These visualizations helped identify patterns and outliers in the data, such as:
   - **Premium and Claim Distributions:** Visualizations like histograms and box plots were used to understand the spread of premiums and claims.
   - **Correlations:** Heatmaps were used to explore the correlation between different variables, such as premiums and claims.
4. **Machine Learning Implementation:** In this project, machine learning techniques were employed to predict future claims or identify risk factors for higher claims. Popular algorithms like **logistic regression**, **random forests**, and **decision trees** were considered for classification or regression tasks. The steps followed included:

- **Data Splitting:** The data was split into training and testing sets.
- **Model Training:** Machine learning models were trained using features such as age, gender, policy type, and claim history.
- **Model Evaluation:** The models were evaluated based on accuracy, precision, recall, and other relevant metrics.

The goal was to predict claim amounts or the likelihood of a policyholder filing a claim in the future, enabling the insurance company to make more informed decisions regarding premium adjustments and risk management.

SQL Implementation

SQL Implementation:

SQL was used to manage and query the health insurance data stored in a relational database. It facilitated efficient data retrieval and analysis, making it easier to work with large datasets.

1. **Data Storage:** The health insurance data was stored in a relational database with multiple tables, such as:
    - **Policyholders:** Contained information on policyholder demographics (e.g., ID, name, age, gender).
    - **Policies:** Contained details about the insurance policies (e.g., policy ID, policy type, premium amount, coverage type).
    - **Claims:** Contained information about claims made by policyholders (e.g., claim ID, claim amount, claim date).
2. **Data Retrieval:** SQL was used to query the database for specific data points needed for analysis. Common SQL queries included:
    - **SELECT:** Used to retrieve specific columns (e.g., age, claim amount) from tables.
    - **JOIN:** Used to combine data from multiple tables, such as joining the **Policyholders** and **Claims** tables to analyze claims by age group.
    - **GROUP BY:** Used to aggregate data by specific categories (e.g., total premiums by policy type).

    - **WHERE:** Used to filter data based on specific conditions (e.g., only selecting claims greater than a certain amount).

3. **Aggregations and Calculations:** SQL was essential for performing aggregations and calculations directly in the database. Some examples include:
    - **SUM:** To calculate total premiums or total claims.
    - **AVG:** To calculate the average claim amount.
    - **COUNT:** To count the number of policies or claims in a given period.
    - **CASE Statements:** Used to create custom categories or conditional calculations (e.g., categorizing premium amounts into "low," "medium," or "high" tiers).
4. **Performance Optimization:** As the dataset could be large, SQL queries were optimized for performance by using:
    - **Indexes:** To speed up search operations.
    - **Query Optimization:** Writing efficient queries to minimize processing time and ensure scalability as the data grows.
5. **Reporting and Insights:** SQL queries were also used to generate insights for reporting purposes. For example:
    - **Claim-to-Premium Ratio:** Calculated using SQL to understand the financial sustainability of the insurance company.
    - **Policyholder Demographics:** Used to analyze claim behavior based on policyholder characteristics such as age, gender, and policy type.

Integration of Python, Machine Learning, and SQL

In this project, Python, machine learning, and SQL worked in tandem to provide a comprehensive analysis:

- **Data Extraction:** SQL queries were used to extract the required data from the database.
- **Data Processing:** Python was used to clean and process the extracted data.
- **Modeling:** Machine learning models were developed using Python to predict future claims and risks.

- **Reporting:** Power BI was used to visualize insights derived from the SQL database and Python analysis.

By integrating Python for advanced analysis and machine learning, SQL for efficient data retrieval, and Power BI for visualization, the health insurance project was able to provide actionable insights to optimize premium pricing and claims management.

6. Dashboard Design (Power BI)

The Power BI dashboard is designed to provide a clear overview of the health insurance metrics. It includes:

- **Layout:** The dashboard has a clean, simple layout with filters at the top for policy type, age group, and gender.
- **Features:** Interactive elements such as slicers allow users to drill down into specific segments. There is a bar chart for the total premium versus claims by policy type, and a pie chart for policy renewal rates.
- **Interactivity:** Users can filter the data by policyholder demographics to explore trends and insights more deeply.

---

7. Results and Insights

- **Claim-to-Premium Ratio:** A ratio of 0.85 indicates that claims are near to the premiums collected, which could suggest a need to increase premiums or refine claims management.
- **High Renewal Rate:** The renewal rate was 75%, indicating a loyal customer base. Further marketing efforts can target the 25% non-renewed policies.
- **Claims by Age Group:** Older policyholders tend to have higher claims, suggesting a need for age-based premium adjustments.
- **Geographical Insights:** Premiums and claims varied across regions, with higher claims in urban areas.

---

8. Conclusion

The project successfully highlighted key trends in the health insurance data. Recommendations include:

- **Adjust Premiums:** Increase premiums for high-risk groups to balance the claim-to-premium ratio.
- **Focus on Retention:** Enhance efforts to improve the policy renewal rate, especially targeting the non-renewing segment.
- **Age-Based Pricing:** Consider introducing age-based premium plans to align premiums with the higher risks of older age groups.

---

9. References

- Python Libraries: pandas, numpy :- [Health Insurance pdf.pdf]
- SQL Database Documentation: [Health Insurance SQL Documentation.pdf]
- Power BI Tutorials: [Health Insurance Dashboard.pdf]

This document provides a step-by-step guide to the health insurance project, covering the processes from data import and cleaning to analysis and reporting using Excel, Power BI, SQL, and Python.

Insights and Findings:

- ➢ The analysis revealed that smokers tend to incur higher healthcare costs, and there is a strong positive correlation between age and healthcare claims.
- ➢ Individuals with higher BMI were also found to have higher claims due to the associated health risks.
- ➢ Regional variations in claims were observed, but they had less impact compared to factors like smoking and BMI.
- ➢ Gender had minimal influence on the variation in claim amounts.