Assignment: Predicting Future Insurance Claim Amounts

1. Overview & Business Context

Passportcard provides medical insurance for customers in relocation abroad. As part of the service the company pays for medical expenses of various types of services, each such request for payment is termed claim.
Your goal is to predict the total claim amount per customer for the next six months, given their claim history. Our company uses historical data on insurance claims alongside detailed member profiles to manage risk, set premiums, and plan financial reserves. Your analysis will help us identify risk factors and forecast future payouts.

2. Data Description

You will work with two datasets:

- **claims_data:**

Contains detailed records of past insurance claims. Key columns include:

  - **ClaimNumber:** Unique claim identifiers, pay attention there may be several claims per member.
  - **TotPaymentUSD:** Payment amount for each claim (note that negative values may indicate adjustments or refunds)
  - **ServiceDate & PayDate:** Dates when the service was provided and when payment was made
  - **ServiceGroup / ServiceType** Service categorizations (e.g., "Office Visit," "Physical Therapy," "Emergency Services")
  - **Member_ID:** The key field for linking to member profiles
  - **PolicyID**: unique identifier of member insurance policy. Note that several members can be linked to the same policy, for example, family members may be under the same policy.
  - **Additional fields:** Include procedure codes, diagnosis indicators, and cost flags (such as maternity_cost, cancer_indicator, inpatient_cost, etc.) that offer insights into the nature of the claims.

- **member_data:**

Contains customer and policy-related information with the following sample columns:

  - **PolicyID, Member_ID:** Identification fields, corresponding to claims_data
  - **PolicyStartDate, PolicyEndDate:** Start and end of customer policy. Note that end date may be in the future indicating that the policy will be active until that date.
  - **DateOfBirth:** date of birth of the member
  - **CountryOfOrigin** - member country of origin
  - **CountryOfDestination** - country where the member resides and gets his medical care
  - **Gender** - True is male

- **Multiple Questionnaire Fields:** (e.g., Questionnaire_hiv, Questionnaire_smoke, Questionnaire_cancer, etc.) that capture health risk factors and lifestyle attributes. During initiation of a policy, the customer answers medical questioners, 1 indicates that the member answered yes to the questionnaire in the column.
- **uw_pct, BMI, average_us, average_ov, average_ob:** Numeric metrics that may help quantify risk.

Note: The common field between these datasets is **Member_ID**. It is essential to join these datasets accurately to leverage the complete information available for each customer.

3. Task Objectives

1. **Data Exploration & Cleaning:**
    - **Understand the Data:**

Review both datasets to grasp variable types, distributions, and potential anomalies.

- **Quality Checks:**
    - Address missing or inconsistent values in both claims and member data.
    - Consider how negative values in TotPaymentUSD should be interpreted (e.g., adjustments, refunds).

1. **Feature Engineering:**
    - **Temporal and Aggregation Features:**
        - Consider available temporal information, for example is there a differences between members with 5-year policy history and those with 1-year history. How would you incorporate this difference into the model?
        - Consider how you'd aggregate historical claim data per customer
    - **Service and Diagnosis Details:**
        - Address the given information regarding the type of service in your model
    - **Customer Profile Insights:**
        - Address demographic and questioner data in your model.
    - **Combining Insights:**

Suggest features that combine the above insights. Consider whether you have enough data for the suggested number of features.

Modeling:

- **Model Selection:**

Develop a model that predicts the total claim amount for the next six months for each customer. You may utilize a simple model to save time and discuss how you'd improve it in a real-life scenario

- **Training:** Describe your training strategy to get the best result and mediate overfitting.

- **Evaluation:**
    - **Performance Metrics:** explain your chosen evaluation metrics

    - **Validation:** Suggest a validation strategy.
    - **Interpretability:** Suggest a way to interpret the results and how you'd make the model interpretable for business stakeholders that must make decisions based on the model. What kind of decisions can and can't be made based on this model?
- **Deliverables:**
    - A well-documented analysis Jupyter Notebook that includes:
        - Data exploration and cleaning steps
        - Feature engineering and transformation processes
        - Model development and evaluation steps
    - Within the notebook address the following:
        - Your analytical approach and key findings
        - The assumptions made during the analysis
        - Potential business actions based on your insights
    - Visualizations (charts/graphs) to support your narrative and illustrate key points.
    - If you don't have enough time to implement something, describe in text what would you do given more time.

4. Additional Instructions

- **Time Frame:**

The task should be designed to be completed within two days.