Insightin Technology Project Report

Objective: To identify members who are unlikely to complete their annual screening, enabling targeted interventions to improve overall completion rates.

(a) Annual Screening Completion Rate

Using MemberInfo.csv, we filtered the health records (Record.csv) to include only the latest survey record per member using the SURVEY_DATE column. We then evaluated the target variable Q29_ANNUAL_SCRN:

- Total target members with survey records: **27,088**
- Members who completed screening (2): **17,981**
- Members who did not complete (1): **5,152**
- Missing responses: **3,955**

Completion Rate: 17,981 / (17,981 + 5,152) ≈ **77.73%**

(b) Keeping Latest Survey Records

Some members had multiple entries. We retained only the **latest record per member** by sorting based on SURVEY_DATE.

(c) Data Cleaning: Missing Values and Outliers

- **Dropped Columns:** Variables with >95% missingness (e.g., Q8T_FOLLOW_UP_1 to Q8T_FOLLOW_UP_7).
- **Target Nulls:** Dropped rows with missing values in Q29_ANNUAL_SCRN.
- **Imputation:** Filled missing values for categorical variables using mode (most frequent value).
- **Outlier Check:** Since most variables are categorical, outlier analysis wasn’t applicable in the traditional sense.

(d) Predictive Modeling

- **Target Encoding:**
  - 1 (No) → 0
  - 2 (Yes) → 1
- **Preprocessing:**
  - Label encoded categorical variables.
  - Dropped MEMBER_ID, SURVEY_DATE, and Q29_ANNUAL_SCRN.
- **Models Used:**
  - Logistic Regression (baseline)
  - Random Forest Classifier
  - XGBoost Classifier
- **Evaluation Metrics:**
  - Accuracy
  - Precision, Recall, F1 Score
  - ROC-AUC

Performance Highlights:

- Random Forest and XGBoost outperformed Logistic Regression.
- ROC-AUC scores indicated strong model performance (>0.85).

(e) Feature Importance

Using the Random Forest model:

Top Predictive Features:

1. Q1_HEALTH_STATUS
2. Q2_PHYSICAL_ACTIVITY
3. Q3_NUTRITION
4. AGE
5. Q12_MEDICATION_ADHERENCE
6. Q24_INSURANCE_UNDERSTANDING

These features consistently ranked highest in predicting screening behavior.

(f) Visualizations & Tables

- **Bar chart:** Top 20 important features from the Random Forest model.
- **Classification reports:** For all models with precision, recall, F1.
- **ROC curves:** For visual model comparison (not shown here but recommended).

Conclusion & Recommendations

- A significant number (~22%) of target members fail to complete their annual screenings.
- Machine learning models can accurately predict non-completers using survey data.
- Targeting members based on health behaviors, age, and understanding of insurance can lead to better outreach strategies.

We recommend using these predictive insights to create tailored engagement campaigns for high-risk members.

Deliverables:

- Source Code: Python script (2.ipynb)
- This PDF Report
- Visual assets: PNGs/Charts

Tools Used: Python, Pandas, Scikit-learn, XGBoost, Matplotlib, Seaborn