Data-Centric AI and Data Profiling - Health Insurance Dataset

Francisco da Ana (up202108762)
(Dated: October 20, 2024)

This report details a data profiling analysis of a health insurance dataset. The analysis uncovers key characteristics, data quality issues and patterns in the dataset, which can have significant implications for future data modeling.

I. INTRODUCTION

The dataset used in this analysis contains 72,458 customer records, across 15 features that describe various socio-economic and demographic characteristics. The target variable is whether or not the customer has **health insurance**, which presents opportunities for classification tasks in predictive modeling.

II. DATA OVERVIEW AND STRUCTURE

Table I provides a summary of the feature types and key properties. There are 15 columns, with a mix of numerical, categorical, and boolean data types.

Feature	Data Type
unnamed	Numeric
custid	Text
age	Numeric
sex	Categorical
income	Numeric
health_ins	Boolean
num_vehicles	Numeric
marital_status	Categorical
housing_type	Categorical
is_employed	Boolean
state_of_res	Text
code_column	Text
gas_usage	Numeric
rooms	Numeric
recent_move	Boolean

TABLE I. Dataset Features and Data Types

Overall, the dataset offers a diverse set of features, each of which plays a role in determining whether a customer is likely to have health insurance. However, there are also significant issues with missing values, outliers, and imbalance in the target variable.

III. DATA PROFILING SUMMARY

Each feature has unique characteristics, which could impact analysis and model performance. This topic will explore a more detailed observation of some variables.

1. Unique identifiers

Unnamed and **Custid**: these features indicate a unique value for each record. The `unnamed` column is likely a numeric index column, while the `custid` is the unique identifier of the costumer using a different format. Due to their nature, these features would not be useful in a classification problem.

2. Age and Income

Age: The age feature contains some anomalous values, including entries for customers aged 0 and above 100. These values are likely data entry errors and should be addressed through data cleaning.

Income: The income distribution is also problematic, with several negative values indicating data errors. The range extends from -6,900 to over 1.2 million, which suggests potential inaccuracies or outliers that will skew analysis.

3. Categorical Features

Sex and **Marital Status**: The sex feature is balanced between males and females, while the marital status feature shows a strong predominance of married individuals (52.5% of the records). Both of these features have no missing values, which makes them easier to incorporate into models.

Employment Status and **Health Insurance**: The `is_employed` feature reveals interesting insights. Among the non-missing values, 95% of the customers are employed, and 90% of them have health insurance. Interestingly, for the unemployed customers, 75% still have health insurance. This suggests that even those without employment maintain health insurance coverage.

IV. KEY DATA QUALITY ISSUES AND ANALYSIS

A. Unique Identifiers

Issue: Both *Unnamed* (a numeric index) and Custid (a customer ID) provide unique values for each record, but do not add value to classification tasks as they don't

contribute to identifying patterns. **Consequences:** Including them could introduce unnecessary noise and lead to overfitting since they don't hold meaningful information for prediction. **Solution:** These columns should be excluded from the model to avoid confusion and improve performance.

B. Missing Values in Employment Status

Issue: The `is_employed` column has 32% missing data, which could severely impact any analysis related to employment. **Consequences:** High levels of missing data in this key feature could lead to biased results or reduced model accuracy, particularly if employment status is a significant predictor for health insurance. **Solutions:** Possible solutions include imputing missing values using the most frequent value or creating a new category to represent missing data.

C. Outliers in Age and Income

Issue: Both the `age` and `income` features have extreme values that are likely erroneous. These outliers could distort statistical summaries and models that rely on these features. **Consequences:** If not handled, these outliers will introduce skewness in regression models and mislead any classification efforts based on income. **Solutions:** A reasonable approach would be to remove or cap outliers beyond certain thresholds. For example, ages below 20 and above 100 could be excluded, and negative income values should be corrected or removed.

D. Imbalanced Target Variable

Issue: The target variable (`health_ins`) is highly imbalanced, with 90.5% of the customers having health insurance. **Consequences:** Models trained on this data will likely favor the majority class, leading to poor performance in predicting the minority class (i.e., customers without health insurance). **Solutions:** Resampling techniques, such as oversampling the minority class or using cost-sensitive learning approaches, can help balance the model's performance.

V. CORRELATION AND FEATURE RELATIONSHIPS

A correlation matrix was generated to understand relationships between the features. Although most features show weak correlations, there are some notable patterns:

- A strong correlation between `marital status` and `age`, indicating older individuals are more likely to be married or widowed.
- A moderate correlation between `housing type` and `recent move`, with renters being more likely to have moved recently.

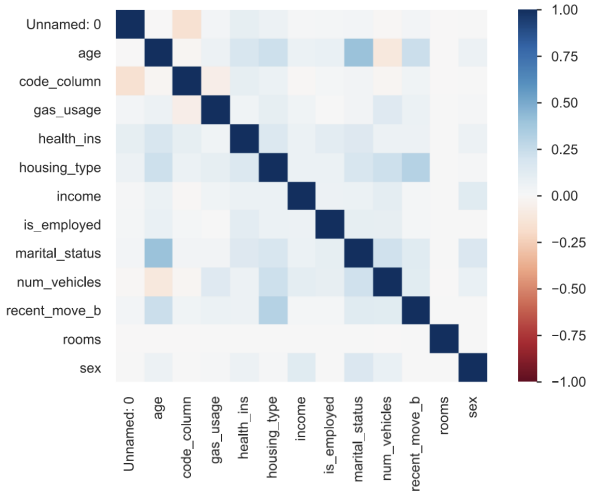

FIG. 1. Correlation Matrix

VI. CONCLUSIONS AND RECOMMENDATIONS

In conclusion, the data profiling exercise has highlighted several important aspects of the health insurance dataset. Key issues include missing data, outliers, and class imbalance, all of which could negatively impact model performance if left unaddressed. Future work should focus on:

The proposed improvements in Section IV will enhance the quality of the dataset and ensure more robust predictive modeling in future analyses.

The *Jupyter Notebook* file developed for this assignment contains a more extended observation of the dataset characteristics.

VII. REFERENCES

- The dataset was collected from the materials of the curricular unit Introduction to Data Science (Master Degree in Artificial Intelligence, University of Porto).
- Miriam Seoane Santos, Artificial Intelligence and Society - Data-Centric AI & Data Profiling.