Assignment 2: Data Preprocessing for Case Competition

Instructions

Learn, firsthand, how top teams tackle analytics problems and extract strategies you can reuse in your case competition. Choose one prize-winning report from the list provided (but make sure no one else in your group uses the same report): Focus on the approach, not the application domain, so you can generalize what you learn to your own case.

• Analyze the Data preparation sections of this report (look for sections titled data preprocessing, data cleaning, data transformation, or feature engineering)
 - a. What specific steps did they take to create, remove, or transform variables?
 - b. Why did they use each step? (Provide the rationale.)
 - c. Which of these steps will you try on your dataset, and why?

Report Chosen: Humana-Mays Healthcare Analytics Case Competition (2020)

A. Specific Steps They Took to Create, Remove, or Transform Variables

1. Exploratory Data Analysis

- Dataset structure: The data had 69,572 Medicare members and 826 fields, including consumer data, medical claims, pharmacy claims, lab claims, demographics, credit data, condition-related features, and CMS features.
- Age analysis: The average age was 70.81 years, with most members between 66 and 77 years. Younger members reported more transportation issues than older members.
- Disability patterns: 23.18% of disabled members had transportation issues compared to 12.19% of non-disabled members.
- Health score correlation: Higher health risk scores (CCI, DSCI, FCI, HCC) were linked with more transportation issues.
- Other patterns: Smokers, renters, single-parent households, and high prescription drug users showed higher transportation issues than others.

2. Feature Engineering Techniques

- Group Binning and Ranking: The team grouped categorical variables such as education, household type, homeownership, and language into clearer categories. Numerical values were converted into percentile ranks to show relative standing. This means that instead of saying someone’s income is $40,000, the model could see that the person falls in, for example, the 30th percentile of income. After creating percentile ranks for many different variables, they combined these ranks into scoring metrics.
- Weighted Metrics Creation: The team noticed that some important ideas, like “stress” or “mobility,” couldn’t be explained by a single variable. So instead of relying on one

column, they built composite scores that combined several related features. They built composite scores for StressIndex (demographic + credit) and MobilityIndex (demographic + health). Both were normalized to a 0–100 scale so they were easy to interpret and compare.

- K-Means Clustering: Used percentile ranks and weighted metrics to create clusters of members with similar characteristics. To make clustering fair, all variables were standardized so they were on the same scale. This process created 30 new cluster-based features. Each cluster grouped members by credit, health, stress, or age patterns.
- Isolation Forest for Anomaly Detection: Outliers can distort model performance, so the team used the Isolation Forest algorithm to identify members whose data looked very different from the majority. They assigned anomaly scores and created a binary flag for outliers. About 5,600 anomalies were identified, and they had a much higher rate of transportation issues (29.54% vs. 13.3%).

3. Feature Selection Process

After feature engineering and dimensionality reduction, the dataset still had thousands of features. Using all of them would risk overfitting (the model learning noise instead of patterns) and would also increase computation time.

- Step 1: Forward selection with Logistic Regression, they added features one by one to a logistic regression model, keeping those that improved the ROC AUC score..
- Step 2: They trained a Random Forest model and used its internal feature importance scores to keep about 250 of the most useful features
- Step 3: From the 250 features, they again used forward selection, this time with the XGBoost algorithm, to choose the final 74 features that gave the best improvement in predictive performance.

B. Rationale for Each Step

1. Exploratory Data Analysis (EDA) Rationale

- EDA was important with 800+ features to get a clear picture of the data before making new features.
- EDA also showed useful domain patterns, like how age, disability, and prescription use were linked to transportation issues.

2. Feature Engineering Rationale

- Group Binning and Ranking: Made raw data easier to work with, preserved order in numbers, and built simple scoring metrics.
- Weighted Metrics: Combined data from different sources into stress and mobility scores, scaled 0–100 for clarity and easy use.
- K-Means Clustering: Found hidden groups of members, with standardization making distances fair and accurate.
- Isolation Forest: Flagged unusual members who had twice the transportation issue rate of normal members, and added simple binary flags for clarity.

3. Feature Selection Rationale

- Reducing 8,000+ features to 74 avoided overfitting and made the model easier to handle.
- Using Logistic Regression, Random Forest, and XGBoost together gave a more balanced view of which features mattered most.
- ROC AUC was used throughout to make sure features improved prediction quality.
- The final 74 features gave a good balance of performance, efficiency, and clarity.

C. Steps to Apply to Our Dataset and Why

1. Systematic Exploratory Data Analysis

We will begin by reviewing the dataset shape, target distribution, and basic descriptive statistics. Like their disability and age analysis, we will examine default rates by age, income groups, and gender to reveal useful patterns.

2. Group Binning and Percentile Ranking

We will bin categories like education into Basic, Intermediate, and Advanced groups. We will also convert key numeric variables such as income, credit, and annuity into percentile ranks. Finally, we will combine these into a Financial Profile Score that summarizes overall financial standing.

3. K-Means Clustering for Customer Segmentation

We will apply K-Means clustering on standardized features such as stress, stability, income, age, and credit. This will divide customers into 3–4 clusters, helping us identify risk segments with different default rates.

4. Three-Step Feature Selection Process

We will select features in three stages. First, apply statistical tests to filter. Second, use Random Forest importance to keep the strongest predictors. Third, refine with XGBoost to choose the final 50–75 features. This ensures a strong but efficient model.