Technical Report: Medical Insurance Cost Prediction

Data Analysis & Machine Learning Team

January 22, 2026

Project Overview

This report details the development and evaluation of a predictive model designed to estimate annual medical insurance charges. By analyzing individual demographic and health factors, the model aims to provide accurate financial forecasting for insurance beneficiaries.

Dataset Overview

The analysis utilizes the `insurance.csv` dataset, consisting of **1,338 records**. Each record represents an individual customer with the following attributes:

Feature	Description	Type
Age	Age of the primary beneficiary (18 - 64 years)	Numerical
Sex	Gender of the contractor (female, male)	Categorical
BMI	Body Mass Index (kg/m^2)	Numerical
Children	Number of children/dependents covered	Numerical
Smoker	Smoking status (yes, no)	Categorical
Region	Residential area in the US	Categorical
Charges	**Annual medical costs billed (Target)**	**Numerical**

Table 1: Dataset Feature Dictionary

Exploratory Data Analysis (EDA)

Our diagnostic checks revealed several key patterns in the data:

- **Univariate Analysis:** The distribution of `Charges` is significantly **right-skewed**. Most individuals incur costs between $5,000 and $12,000, while high-risk outliers exceed $50,000.
- **Bivariate Analysis:**
  - **Smoker vs. Charges:** This is the strongest correlation. Smokers incur vastly higher costs compared to non-smokers.
  - **Age & BMI vs. Charges:** Both variables show a positive linear trend with respect to costs.
- **Correlation Analysis:** Heatmap visualization confirmed that `Smoker`, `Age`, and `BMI` are the primary drivers of insurance pricing.

Data Preprocessing

To prepare the data for the Linear Regression engine, the following steps were implemented:

1. **Standardization:** Input strings were processed using `.lower()` and `.strip()` to maintain case-insensitive consistency.
2. **Encoding:**
    - `Sex` and `Smoker` were binary encoded (0/1).
    - `Region` was processed via One-Hot Encoding (`pd.get_dummies`) with `drop_first=True` to avoid the dummy variable trap.
3. **Feature Matrix:** Data was split into features (X) and target (y) with an 80/20 train-test ratio.

Model Performance & Limitations

We utilized a **Linear Regression** model to find the line of best fit.

Evaluation Metrics

- **R-Squared (R^2):** ≈ 0.75. The model explains 75% of the variance in medical costs.
- **Mean Absolute Error (MAE):** $\approx \$4,100$.

Model Flaws & Technical Challenges

Despite a strong R^2 score, the model exhibits specific weaknesses:

- **Average Prediction Error:** An MAE of $4,100 is significant. For a $10,000 prediction, the actual cost may range from $6,000 to $14,000.
- **Outlier Sensitivity:** Linear models struggle with extreme medical events (outliers) that result in charges over $50,000.
- **Non-Linear Patterns:** The relationship between BMI and Charges becomes non-linear (exponential risk) for individuals with a BMI over 30, which the current model underestimates.

Conclusion

The Linear Regression model serves as a robust baseline for insurance cost estimation. However, to address the $4,100 error margin and better handle outliers, we recommend exploring non-linear algorithms such as **Random Forest Regressor** or **XGBoost** in the next phase of development.