**UC BERKELEY PROFESSIONAL CERTIFICATE IN MACHINE LEARNING AND AI**

**FINAL CAPSTONE PROJECT**

**[HEALTH INSURANCE COST PREDICTION & RISK ANALYSIS](https://github.com/kfmatovic716/HEALTH-INSURANCE-COST-ML.git)**

**[USING MACHINE LEARNING TECHNIQUES](https://github.com/kfmatovic716/HEALTH-INSURANCE-COST-ML.git)**

*By Katherine Matovic*

**INTRODUCTION**

The growing emphasis of health equity has recently become a major priority in the United States, and it has contributed to the rising healthcare costs. Leveraging health insurance data to predict future claims costs and identifying key factors that drive these escalating costs can provide actionable insights to improve decision-making and expense management of health insurance providers.

**PROBLEM STATEMENTS**

1. Based on an individual’s demographic and lifestyle, can we predict the individual’s health insurance charges accurately?
2. Can we classify individuals who are likely to incur **high** versus **low** health insurance costs?
3. What factors strongly influence health insurance costs?

**GOALS**

1. **Cost Prediction** – Build a baseline predictive model using **Linear Regression**, then enhance performance using **Ridge and Lasso Regression**. Evaluate models using **R-squared, MAE, and MSE** metrics.
2. **Risk Classification** – Classify individuals into **high-cos**t **or low-cost categories** using **Logistic Regression**, based on lifestyle features and/or predicted charges.
3. **Feature Importance** – Identify the most influential factors driving insurance costs using **GridSearchCV** for model optimization. (age and smoker)

**DATA ACQUISITION AND DESCRIPTION**

* **SOURCE:** [Kaggle – Medical Insurance Cost Prediction](https://www.kaggle.com/code/alyashoush/medical-insurance-cost-prediction/input)
* **DATA SIZE:** The raw data has 7 features and 2,772 records in total
* **DESCRIPTION**:

The dataset has the following features and their descriptions:

* age - person’s age
* sex - person’s gender
* bmi - body mass index
* children – policyholder’s number of children
* smoker - indicates if person is smoker (“yes”) or non-smoker (“no”)
* region - US region where person resides
* charges (target variable) - insurance premium price

**DATA TRANSFORMATIONS**

* Utilized the map function by converting all categorical variables (sex, region, and smoker) into a numeric form. This allows them to be included in the correlation matrix to explore relationships with other features and supports the modeling stage by ensuring compatibility with running machine learning algorithms.
* Utilized Standard Scaler to scale final dataset to be used in modeling before training data

**EXPLORATORY DATA ANALYSIS (EDA)**

1. **Missing Values and Duplicate Records**

* There were no missing values in the column features
* There were 1,435 duplicate rows that were deleted from the raw data; Some of the records didn't just have one duplicate but 3 duplicates!
* There were total of 1,337 records in the new dataset after deleting duplicates

1. **Univariate Analysis and Outlier Detection**

**FEATURE 1: AGE**

**A graph with a number of blue and green lines

AI-generated content may be incorrect.**

* The age distribution is skewed to the right, which means that many individuals fall within the younger age range (18–20). The curve then stretches gradually toward older ages, extending up to age 63. The smallest representation is seen at ages 64–65, which aligns with expectations in the context of insurance: individuals over 65 are typically considered higher risk, making them fewer desirable participants for many insurance plans.
* Age appears to have a consistent distribution and no present outliers in the feature

**FEATURE 2: BODY MASS INDEX (BMI)**

A diagram of a mass index distribution

AI-generated content may be incorrect.

* BMI categories are as follows: underweight are <18.5, healthy weight: 18.5 to less than 25, overweight: 25.0 to less than 30.0 and obesity: >=30.0
* The BMI illustrates an approximately normal distribution, with the highest concentration of values between 25 and 35 which corresponds to the overweight and obese categories. The tail-end distributions represent individuals in the healthy weight range on the lower end and the severely obese group on the upper end.
* This distribution reflects prevailing health and lifestyle patterns in the US. A significant proportion of the population is overweight, largely attributed to poor nutrition, high stress levels, and sedentary lifestyles.
* There were a few outliers (approximately 7) with BMI over 47 that are severely obese

**FEATURE 3: PARTICIPANT’S NUMBER OF CHILDREN**

A graph with a number of children

AI-generated content may be incorrect.

* The 'number of children’ feature shows a right-skewed distribution. Most participants (approximately 575) reported having no children, followed by 324 participants with one child. The frequency declines progressively as the number of children increases, forming a long tail toward higher values.
* There were no outliers to be observed in this feature

**FEATURE 4: HEALTH INSURANCE PREMIUM COST (TARGET)**

A graph of a cost distribution

AI-generated content may be incorrect.

* The health insurance charges, which serves as the target variable, is right-skewed, indicating that much of the population pays less than $14K. However, there is a notable presence of outliers, with some individuals incurring costs exceeding $34K.
* The dataset has participants who have high premium charges, which are considered outliers. It is important to keep them because they usually reflect real people with unusual but valid health/risk profiles — i.e., older policyholders, smokers, or those with chronic conditions. A biased model may result and distort the true distribution of costs when excluded.
* Outliers are critical to model generalization especially with a classification task that identifies policyholders who are low and high risk. Eliminating these records could make the model less effective in recognizing the most important class.
* Outliers play a crucial role in business decision making, particularly since in this dataset they mostly represent high-premium policyholders (with costs above $34K). These cases not only influence premium pricing strategies but also help insurers anticipate and prepare for costly claims

**FEATURE 5: GENDER**

A diagram of a distribution

AI-generated content may be incorrect.

* Population in gender feature is evenly distributed, approximately 50% are males and 50% female participants, which reduces bias in model training
* Regulators emphasize non-discriminatory models especially in sensitive domains like healthcare and insurance. A balanced distribution meets this ethical consideration and fairness requirements

**FEATURE 6: SMOKER STATUS**

**A green and pink circle with a number of percentages

AI-generated content may be incorrect.**

* There is approximately 80% non-smoker population in this dataset, indicating class imbalance.
* The model might achieve high accuracy by mostly predicting “non-smoker,” but it would perform poorly in detecting smokers without adjustment

**FEATURE 7: REGION**

**A graph of a distribution

AI-generated content may be incorrect.**

* The regions of origin are evenly distributed among participants in this dataset, except for the Southeast, which has a slightly higher representation

1. **Bivariate Analysis**

**FIGURE1. CORRELATION MATRIX**

A red and blue squares with white text

AI-generated content may be incorrect.

* Based on the correlation matrix, the top 3 features that have positive relationship with healthcare insurance premium charges are: smoker (~80% correlation), age (~30%) and BMI (~20%).
* Having children, gender and the region a policyholder is coming from are features that does not directly impact healthcare insurance charges

**FIGURE2. HEALH INSURANCE PREMIUM CHARGES VS SMOKER**

**A graph showing a comparison of smoking status

AI-generated content may be incorrect.**

* The feature most strongly associated with health insurance costs is smoking status of a policyholder. This aligns with real-world evidence, as smokers face a significantly higher risk of developing life-threatening conditions such as cancer, stroke, and diabetes. As a result, they are classified as high-risk policyholders, making smoking status a critical factor in predicting insurance premiums.
* The figure shows that smokers exhibit much greater variability in insurance premium costs, with charges reaching up to $60K, compared to non-smokers whose costs peak around $20K. The difference is significant, with smokers having a median cost of approximately $34K versus just $7K for non-smokers.
* The non-smokers have some outliers that are not much of a significant concern because they are still lower than the typical cost for smokers

**FIGURE3. HEALH INSURANCE PREMIUM CHARGES VS AGE**

**A graph showing the number of insurance charges

AI-generated content may be incorrect.**

* The red regression line illustrates a positive correlation between the insurance charges and the policyholder's age, and insurance tend to increase as the policyholder gets older**.**
* At every age level, insurance costs vary widely, ranging from about $1,000 to $60,000. These variations are likely driven by additional factors such as smoking status, chronic health conditions, and BMI. In general, younger policyholders (age < 35) have lower charges but some charges were high mainly due to some outliers, who might be a smoker, has high BMI or has chronic disease.
* Although the trendline shows a steady increase with age, cost variations remain large among older policyholders (age > 50), indicating that not all older individuals pay high insurance premiums.

**FIGURE4. HEALH INSURANCE PREMIUM CHARGES VS BMI**

A diagram of a graph showing a number of insurance charges

AI-generated content may be incorrect.

* The regression line shows an upward trend and a shallow slope. This suggests that higher BMI is associated with higher insurance charges but does not appear to be a strong predictor of charges due to the shallow slope.
* Just like the relationship between age and insurance charges above, insurance charges vary widely at every BMI value, ranging from approximately $1,000 to over $60,000. It is likely that other factors such as smoker status, age and chronic diseases play a much significant role in determining costs.
* Most policyholders fall in between approximately 20 to 35, where charges vary significantly. This suggests that showing that BMI is not a reliable standalone predictor of charges.
* Higher charges above $40K appear more frequently in BMI above 30. On the other hand, some with very high BMI are still in low charges. This again indicates that there are other factors that interacts with BMI in predicting charges.

1. **Multivariate Analysis**
2. **FIGURE5. PAIRPLOT**

**A graph of health insurance features

AI-generated content may be incorrect.**

* Across all ages, smokers consistently incur higher charges than non-smokers. In general, older policyholders pay more charges but with smoking status creates a significant difference than age alone.
* Smokers with high BMI over 30 incur high charges (some over $60K). Non-smokers show less of a clear pattern since many with high BMI still fall in the low-to-mid insurance charge range. This means that smoking amplifies the effect of BMI on costs.
* There is no clear trend between age and BMI. In this dataset, age and BMI are independent. Smokers and non-smokers are scattered similarly across the BMI and age spectrum.
* Smoking status is the strongest, most consistent driver of high insurance charges across all features. In almost every subplot, smokers (yellow) are shifted charges upward compared to non-smokers which reinforces the idea that smoking status is a key driver in predicting insurance costs, more influential than age or BMI.

**MODELLING**

**Baseline Model: Linear Regression to Predict Health Insurance Charges**

***Coefficients and Intercept:***

* Each coefficient is the expected dollar change in charges for a +1 unit increase in that feature, holding others constant. It shows how the prediction changes if you change just that one feature, while pretending everything else in the model stays the same. For example, when two people with the same features as age, gender, bmi, and number of children but one is a smoker and the other person is not, the model would predict a premium for the smoker to be $23k higher.
* The ***intercept*** is not much meaningful to predict baseline charges when other features zero. In this insurance dataset, values of features like age, BMI and smoking status are very important in predicting health insurance charges.

|  |  |
| --- | --- |
| **FEATURE** | **COEFFICIENTS** |
| Age | 248.76 |
| Sex | -99.70 |
| BMI | 312.61 |
| Children | 534.12 |
| Smoker Status | 23052.15 |
| Region | 237.63 |

***Evaluation Metrics***:

* The model illustrates approximately 81% of the variation in insurance charges across individuals. The differences in charges between policyholders can be accounted for by the features such as age, sex, BMI, children, and smoking status.
* Predictions on average are off by approximately $6k (RMSE test). For example, if the true insurance charge for a policyholder is $4,000, the model will predict it as $10k. The model usually predicts within about $4k (MAE test) of the true value. Occasionally it misses by much more (like $10k+), which nudges the RMSE up to $6k. Since RMSE isn’t significantly different with MAE, those bigger errors are present but not overwhelming.

|  |  |  |
| --- | --- | --- |
| **METRICS** | **TRAIN** | **TEST** |
| R2 Score | 0.7297 | 0.8068 |
| RMSE | 6083.2172 | 5957.6088 |
| MAE | 4181.3216 | 4182.3532 |

***Feature Importance:***

A blue and white rectangle with black border

AI-generated content may be incorrect.