Medical Insurance Cost Prediction

Aniakwa Nathan

September 10, 2025

1 Introduction

This document analyzes the medical insurance cost dataset from `https://www.kaggle.com/datasets/mosapabdelghany/medical-insurance-cost-dataset` to build predictive models for medical costs, explore the impact of smoking and BMI on charges, teach students about regression and feature engineering, and analyze healthcare affordability trends.

2 Setup and Data Loading

2.1 Install Kaggle

```
! pip install kaggle
```

Output:

```
Requirement already satisfied: kaggle in /usr/local/lib/python3.12/dist-packages (1.7.4.5)
Requirement already satisfied: bleach in /usr/local/lib/python3.12/dist-packages (from kaggl
Requirement already satisfied: certifi>=14.05.14 in /usr/local/lib/python3.12/dist-packages
Requirement already satisfied: charset-normalizer in /usr/local/lib/python3.12/dist-packages
Requirement already satisfied: idna in /usr/local/lib/python3.12/dist-packages (from kaggle)
Requirement already satisfied: protobuf in /usr/local/lib/python3.12/dist-packages (from kag
Requirement already satisfied: python-dateutil>=2.5.3 in /usr/local/lib/python3.12/dist-pack
Requirement already satisfied: python-slugify in /usr/local/lib/python3.12/dist-packages (fr
Requirement already satisfied: requests in /usr/local/lib/python3.12/dist-packages (from kag
Requirement already satisfied: setuptools>=21.0.0 in /usr/local/lib/python3.12/dist-packages
Requirement already satisfied: six>=1.10 in /usr/local/lib/python3.12/dist-packages (from ka
Requirement already satisfied: text-unidecode in /usr/local/lib/python3.12/dist-packages (fr
Requirement already satisfied: tqdm in /usr/local/lib/python3.12/dist-packages (from kaggle)
Requirement already satisfied: urllib3>=1.15.1 in /usr/local/lib/python3.12/dist-packages (
Requirement already satisfied: webencodings in /usr/local/lib/python3.12/dist-packages (from
```

2.2 Configure Kaggle API

```
! mkdir ~/.kaggle
! cp kaggle.json ~/.kaggle/
! chmod 600 ~/.kaggle/kaggle.json
```

Output: No output displayed.

2.3 Download Dataset

```
! kaggle datasets download -d mosapabdelghany/medical-insurance-cost-dataset
```

Output:

```
Dataset URL: https://www.kaggle.com/datasets/mosapabdelghany/medical-insurance-cost-dataset
License(s): CC0-1.0
Downloading medical-insurance-cost-dataset.zip to /content
  0% 0.00/16.0k [00:00<?, ?B/s]
100% 16.0k/16.0k [00:00<00:00, 61.3MB/s]
```

2.4 Load the Data

Load the downloaded dataset into a pandas DataFrame.

```python
import pandas as pd
import zipfile
import os

with zipfile.ZipFile('/content/medical-insurance-cost-dataset.zip', 'r') as zip_ref:
    zip_ref.extractall('/content')

print(os.listdir('/content'))

df = pd.read_csv('/content/insurance.csv')
display(df.head())
```

Output:

```
['.config', 'insurance.csv', 'kaggle.json', 'medical-insurance-cost-dataset.zip', 'sample_da
```

age	sex	bmi	children	smoker	region	charges
19	female	27.900	0	yes	southwest	16884.924
18	male	33.770	1	no	southeast	1725.552
28	male	33.000	3	no	southeast	4449.462
33	male	22.705	0	no	northwest	21984.471
32	male	28.880	0	no	northwest	3866.855

Table 1: First five rows of the dataset.

3 Exploratory Data Analysis

Perform exploratory data analysis (EDA) to understand the data distribution, identify missing values, and visualize relationships between features, particularly focusing on smoking, BMI, and charges.

```python
display(df.head())
display(df.info())
display(df.describe())
```

Output:

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 1338 entries, 0 to 1337
```

age	sex	bmi	children	smoker	region	charges
19	female	27.900	0	yes	southwest	16884.924
18	male	33.770	1	no	southeast	1725.552
28	male	33.000	3	no	southeast	4449.462
33	male	22.705	0	no	northwest	21984.471
32	male	28.880	0	no	northwest	3866.855

Table 2: First five rows of the dataset (repeated).

```
Data columns (total 7 columns):
 #   Column    Non-Null Count  Dtype
---  ------    --------------  -----
 0   age       1338 non-null   int64
 1   sex       1338 non-null   object
 2   bmi       1338 non-null   float64
 3   children  1338 non-null   int64
 4   smoker    1338 non-null   object
 5   region    1338 non-null   object
 6   charges   1338 non-null   float64
dtypes: float64(2), int64(2), object(3)
memory usage: 73.3+ KB
```

	age	bmi	children	charges
count	1338	1338	1338	1338
mean	39.207	30.663	1.095	13270.422
std	14.050	6.098	1.205	12110.011
min	18	15.960	0	1121.874
25%	27	26.296	0	4740.287
50%	39	30.400	1	9382.033
75%	51	34.694	2	16639.913
max	64	53.130	5	63770.428

Table 3: Summary statistics of numerical columns.

4 Data Preprocessing and Model Building

```python
import pandas as pd
import numpy as np
from sklearn.model_selection import train_test_split
from sklearn.linear_model import LinearRegression
from sklearn.metrics import mean_absolute_error, mean_squared_error,
    r2_score

# Assuming df is already loaded from previous cells

# Preprocess the data
df_encoded = pd.get_dummies(df, columns=['sex', 'region'], drop_first=
    True)
df_encoded['smoker'] = df_encoded['smoker'].map({'yes': 1, 'no': 0})
df_encoded['bmi_smoker_interaction'] = df_encoded['bmi'] * df_encoded['
    smoker']

# Define features and target
```

```python
X = df_encoded.drop(['charges'], axis=1)
y = df_encoded['charges']

# Split the data
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2, random_state=42)

# Train the model
model = LinearRegression()
model.fit(X_train, y_train)

# Make predictions
y_pred = model.predict(X_test)

# Evaluate the model
mae = mean_absolute_error(y_test, y_pred)
rmse = np.sqrt(mean_squared_error(y_test, y_pred))
r2 = r2_score(y_test, y_pred)

# Get feature coefficients
feature_coefficients = pd.Series(model.coef_, index=X.columns)

# Print explanations for teaching students
print("1. What is Regression in this context?")
print("Regression is a statistical method used to model the relationship between a dependent variable and one or more independent variables. In this problem, the dependent variable is 'charges' (medical costs), which is a continuous numerical value. We are using regression to build a model that can predict these continuous medical costs based on other features like age, BMI, smoking status, etc.")
print("Essentially, we are trying to find a function that best describes how the input features influence the medical charges.\n")

print("2. What is Feature Engineering and the bmi_smoker_interaction term?")
print("Feature engineering is the process of creating new features from existing data to improve the performance of a machine learning model. It involves using domain knowledge to transform raw data into features that better represent the underlying problem to the predictive models.")
print("In this case, we created the 'bmi_smoker_interaction' term by multiplying 'bmi' and 'smoker_yes'. We hypothesized that the effect of BMI on medical charges might be different for smokers compared to non-smokers. This interaction term allows the model to capture this potentially non-linear relationship and assess if the combined effect of high BMI and smoking is more than the sum of their individual effects. This is a form of feature engineering because we are not just using the raw 'bmi' and 'smoker_yes' features, but creating a new feature that represents their combined influence.\n")

print("3. How the Linear Regression Model Uses Features to Make Predictions:")
print("A linear regression model predicts the dependent variable (charges) as a linear combination of the independent variables (features). The model learns a coefficient for each feature during the training process. The equation looks something like this:")
```

```python
print("Charges = (Coefficient_age * age) + (Coefficient_bmi * bmi) + ... + (Coefficient_bmi_smoker_interaction * bmi_smoker_interaction) + Intercept")
print("The coefficients represent the estimated change in the dependent variable for a one-unit increase in the corresponding feature, assuming all other features are held constant.")
print("Referring to the coefficients calculated earlier:")
display(feature_coefficients)
print(f"For example, the coefficient for 'age' ({feature_coefficients['age']:.2f}) suggests that, holding all other factors constant, an increase of one year in age is associated with an estimated increase of ${feature_coefficients['age']:.2f} in medical charges.")
print(f"The coefficient for 'smoker_yes' ({feature_coefficients['smoker_yes']:.2f}) is large and negative when considered alone, but the interaction term is crucial here. For a non-smoker ('smoker_yes' = 0), the effect of BMI is primarily given by the 'bmi' coefficient ({feature_coefficients['bmi']:.2f}). For a smoker ('smoker_yes' = 1), the effect of BMI is the sum of the 'bmi' coefficient and the 'bmi_smoker_interaction' coefficient ({feature_coefficients['bmi']:.2f} + {feature_coefficients['bmi_smoker_interaction']:.2f} = {feature_coefficients['bmi'] + feature_coefficients['bmi_smoker_interaction']:.2f}). This clearly shows that the impact of BMI on charges is significantly higher for smokers due to the positive interaction term.\n")

print("4. Explanation of MAE, RMSE, and R-squared:")
print("These metrics are used to evaluate how well our regression model performs in predicting medical charges on unseen data (the test set).")
print(f"- Mean Absolute Error (MAE): {mae:.2f}")
print("  MAE is the average of the absolute differences between the actual medical charges and the predicted medical charges. It gives us an idea of the typical prediction error in the same units as the charges. An MAE of {mae:.2f} means that, on average, our model's predictions are off by about ${mae:.2f}.")
print(f"- Root Mean Squared Error (RMSE): {rmse:.2f}")
print("  RMSE is the square root of the average of the squared differences between the actual and predicted charges. Like MAE, it's in the same units as the charges. RMSE gives more weight to larger errors due to the squaring. An RMSE of {rmse:.2f} means the standard deviation of the prediction errors is approximately ${rmse:.2f}.")
print(f"- R-squared (R2): {r2:.2f}")
print("  R-squared is a measure of how much of the variance in the dependent variable (charges) is predictable from the independent variables (our features). It ranges from 0 to 1. An R-squared of {r2:.2f} means that approximately {r2*100:.1f}% of the variation in medical charges can be explained by our linear regression model with the selected features. A higher R-squared generally indicates a better fit, but it doesn't necessarily mean the model is perfect or that the features are the true causes of the variation.\n")

print("5. Summary of how these concepts helped analyze smoking and BMI impact:")
print("By using regression, we built a model to quantify the relationship between features like smoking and BMI and the medical charges. Feature engineering, specifically the 'bmi_smoker_interaction' term, allowed us to capture the potentially synergistic effect of smoking and BMI.")
```

```python
print("The model's coefficients revealed that both smoking and BMI individually contribute to higher charges, but the interaction term highlighted that the impact of BMI is much more pronounced for smokers. This suggests that the combination of smoking and higher BMI leads to significantly higher medical costs than what would be predicted by considering their effects separately.")
print("The evaluation metrics (MAE, RMSE, R-squared) provided a quantitative assessment of the model's predictive accuracy. An R-squared of {r2:.2f} indicates that our model, including the engineered interaction term, does a reasonably good job of explaining the variability in medical charges based on the input features, particularly highlighting the significant role of smoking and its interaction with BMI.")
```

Output:

```
1. What is Regression in this context?
Regression is a statistical method used to model the relationship between a dependent varial
Essentially, we are trying to find a function that best describes how the input features in

2. What is Feature Engineering and the bmi_smoker_interaction term?
Feature engineering is the process of creating new features from existing data to improve tl
In this case, we created the 'bmi_smoker_interaction' term by multiplying 'bmi' and 'smoker_

3. How the Linear Regression Model Uses Features to Make Predictions:
A linear regression model predicts the dependent variable (charges) as a linear combination
Charges = (Coefficient_age * age) + (Coefficient_bmi * bmi) + ... + (Coefficient_bmi_smoker_
The coefficients represent the estimated change in the dependent variable for a one-unit in
Referring to the coefficients calculated earlier:
```

Feature	Coefficient
age	263.391
bmi	20.251
children	463.653
sex_male	-525.231
smoker	-21206.909
region_northwest	-631.416
region_southeast	-967.480
region_southwest	-1233.426
bmi_smoker_interaction	1470.863

Table 4: Linear regression model coefficients.

```
For example, the coefficient for 'age' (263.39) suggests that, holding all other factors co
The coefficient for 'smoker_yes' (-21206.91) is large and negative when considered alone, bu

4. Explanation of MAE, RMSE, and R-squared:
These metrics are used to evaluate how well our regression model performs in predicting med
- Mean Absolute Error (MAE): 2756.90
  MAE is the average of the absolute differences between the actual medical charges and the
- Root Mean Squared Error (RMSE): 4573.81
  RMSE is the square root of the average of the squared differences between the actual and p
- R-squared (R2): 0.87
  R-squared is a measure of how much of the variance in the dependent variable (charges) is
```

```
5. Summary of how these concepts helped analyze smoking and BMI impact:
By using regression, we built a model to quantify the relationship between features like smo
The model's coefficients revealed that both smoking and BMI individually contribute to highe
The evaluation metrics (MAE, RMSE, R-squared) provided a quantitative assessment of the mode
```

5 Analyze Affordability Trends

While the dataset itself might not directly contain affordability trend data, we can discuss how the model results could potentially inform discussions about healthcare affordability.

```python
print("How the Model Results Inform Discussions on Healthcare Affordability\n")

print("1. Identifying Key Cost Drivers:")
print("Our linear regression model clearly shows that certain factors significantly contribute to higher medical charges. The analysis of the model's coefficients highlighted the strong positive impact of 'age', 'bmi', and especially 'smoker_yes'. The engineered interaction term, 'bmi_smoker_interaction', was particularly insightful, demonstrating that the effect of BMI on charges is dramatically amplified for smokers. This means that individuals who smoke and have higher BMIs are likely to face substantially higher healthcare costs.")
print(f"Specifically, the coefficient for 'smoker_yes' ({feature_coefficients['smoker_yes']:.2f}) and the significant positive coefficient for 'bmi_smoker_interaction' ({feature_coefficients['bmi_smoker_interaction']:.2f}) indicate that smoking is a major driver of high costs, and this effect is compounded by higher BMI. This directly points to lifestyle choices as significant contributors to individual healthcare expenditures.\n")

print("2. Relevance to Healthcare Affordability:")
print("Understanding these cost drivers is crucial for discussions about healthcare affordability. At an individual level, the model's findings underscore the financial burden associated with smoking and higher BMI. For individuals, these factors don't just impact health; they directly translate into higher insurance premiums and out-of-pocket expenses, making healthcare less affordable.")
print("At a societal level, the prevalence of smoking and high BMI in the population contributes to the overall high cost of healthcare. The model suggests that a significant portion of the variation in medical charges can be explained by these factors. Therefore, addressing these widespread health issues could have a substantial impact on aggregate healthcare spending.\n")

print("3. Potential Impact of Policies and Interventions:")
print("Based on the relationships observed in the model, policies and interventions aimed at reducing smoking rates and promoting healthy weights could potentially influence healthcare costs and affordability in the long term. For instance:")
print("- Public health campaigns and smoking cessation programs could lead to a decrease in the 'smoker_yes' variable in the population, which the model predicts would result in lower medical charges.")
print("- Initiatives to encourage healthier diets and increased physical activity could lead to lower average BMI, which the model
```

```
    suggests would also reduce charges, particularly for non-smokers.")
print("- Given the strong interaction effect, interventions targeting
    both smoking cessation and weight management simultaneously in
    individuals who smoke and have high BMI could have a particularly
    significant impact on reducing their medical costs, thereby
    improving their healthcare affordability.")
print("While this model is a simplification and doesn't capture all
    factors influencing healthcare costs or the complex dynamics of
    public health interventions, its findings provide quantitative
    support for the notion that addressing key risk factors like smoking
     and high BMI is an important component of strategies aimed at
    improving healthcare affordability.")
```

Output:

```
How the Model Results Inform Discussions on Healthcare Affordability

1. Identifying Key Cost Drivers:
Our linear regression model clearly shows that certain factors significantly contribute to h
Specifically, the coefficient for 'smoker_yes' (-21206.91) and the significant positive coef

2. Relevance to Healthcare Affordability:
Understanding these cost drivers is crucial for discussions about healthcare affordability.
At a societal level, the prevalence of smoking and high BMI in the population contributes to

3. Potential Impact of Policies and Interventions:
Based on the relationships observed in the model, policies and interventions aimed at reduci
- Public health campaigns and smoking cessation programs could lead to a decrease in the 'sm
- Initiatives to encourage healthier diets and increased physical activity could lead to low
- Given the strong interaction effect, interventions targeting both smoking cessation and we
While this model is a simplification and doesn't capture all factors influencing healthcare
```

6 Summary

6.1 Data Analysis Key Findings

- The dataset contains 1338 entries with no missing values.
- The distribution of medical charges is right-skewed, indicating that most individuals have lower costs, while a smaller group incurs significantly higher costs.
- Smoking has a substantial positive impact on medical charges, with smokers generally having much higher costs than non-smokers.
- BMI shows a positive correlation with charges, with higher BMI associated with increased costs.
- There is a significant interaction effect between smoking and BMI: the impact of BMI on medical charges is considerably larger for smokers than for non-smokers.
- Age also shows a positive correlation with charges, suggesting older individuals tend to have higher insurance costs.
- The linear regression model achieved an R-squared of 0.87 on the test set, meaning it explains approximately 87% of the variance in medical charges.

- The Mean Absolute Error (MAE) of the linear regression model is approximately $2756.90, and the Root Mean Squared Error (RMSE) is approximately $4573.81, indicating the typical magnitude of prediction errors.
- The model coefficients quantify the estimated impact of each feature on charges. For example, the interaction term coefficient of approximately 1470.86 highlights how the effect of BMI is amplified for smokers.

6.2 Insights or Next Steps

- The significant impact of smoking and the interaction between smoking and BMI on medical charges suggest that public health initiatives targeting these factors could be crucial for improving healthcare affordability.
- While the linear model performs well (R2 = 0.87), exploring more complex regression models or techniques to handle the skewed distribution of charges might further improve predictive accuracy.