Billy Pierre

Critical Thinking: Linear Regression

24 May 2025

Predicting Medical Insurance Charges Using Linear Regression

Abstract

This project applies linear regression to a real-world insurance dataset to predict medical charges based on demographic and lifestyle factors. The analysis includes data cleaning, encoding, correlation analysis, model training, and evaluation. Results show that smoking status, age, and BMI are the most influential predictors. The final model explains approximately 78% of the variance in charges, with a Mean Squared Error of 33,635,210. Recommendations are made for future model enhancements using non-linear techniques.

Introduction

The rising cost of healthcare has driven interest in predictive modeling to estimate medical expenses. Understanding what factors contribute most to individual insurance charges can aid both insurers and policyholders. This study uses linear regression, a fundamental machine learning technique, to analyze a dataset of medical records and predict charges based on features such as age, sex, body mass index (BMI), number of children, smoking status, and region of residence.

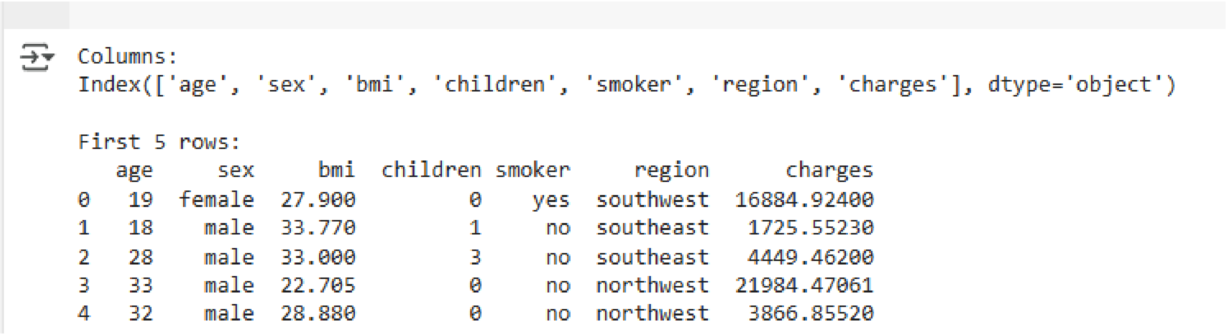
Methods

Dataset Overview

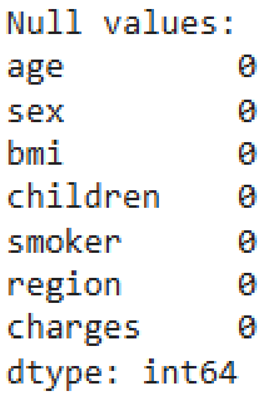
The dataset consists of 1,338 rows and 7 columns:

- `age`: Age of the individual
- `sex`: Gender (encoded)
- `bmi`: Body mass index
- `children`: Number of dependents
- `smoker`: Smoking status (encoded)
- `region`: U.S. region (encoded)
- `charges`: Individual medical insurance cost

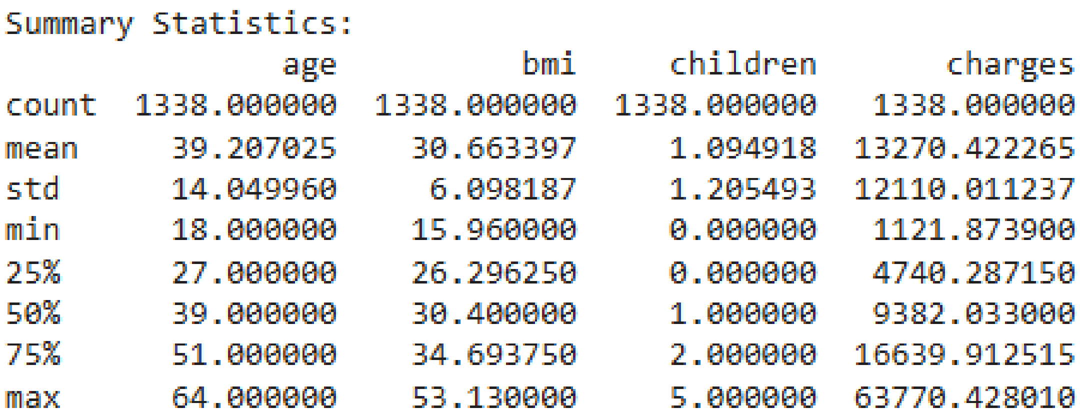
No null values were present in the dataset, and summary statistics are displayed in the table below.



```
Columns:
Index(['age', 'sex', 'bmi', 'children', 'smoker', 'region', 'charges'], dtype='object')

First 5 rows:
   age     sex     bmi  children smoker     region      charges
0   19  female  27.900         0    yes  southwest  16884.92400
1   18    male  33.770         1     no  southeast   1725.55230
2   28    male  33.000         3     no  southeast   4449.46200
3   33    male  22.705         0     no  northwest  21984.47061
4   32    male  28.880         0     no  northwest   3866.85520
```



```
Null values:
age         0
sex         0
bmi         0
children    0
smoker      0
region      0
charges     0
dtype: int64
```



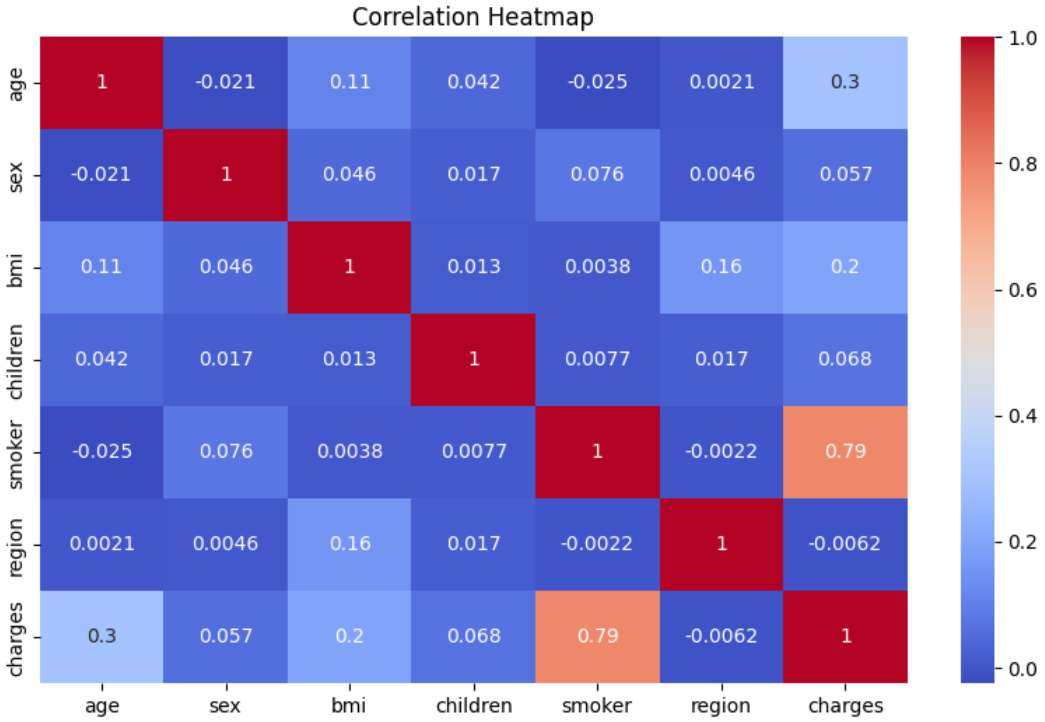
```
Summary Statistics:
               age          bmi     children       charges
count  1338.000000  1338.000000  1338.000000   1338.000000
mean     39.207025    30.663397     1.094918  13270.422265
std      14.049960     6.098187     1.205493  12110.011237
min      18.000000    15.960000     0.000000   1121.873900
25%      27.000000    26.296250     0.000000   4740.287150
50%      39.000000    30.400000     1.000000   9382.033000
75%      51.000000    34.693750     2.000000  16639.912515
max      64.000000    53.130000     5.000000  63770.428010
```

Preprocessing and Encoding

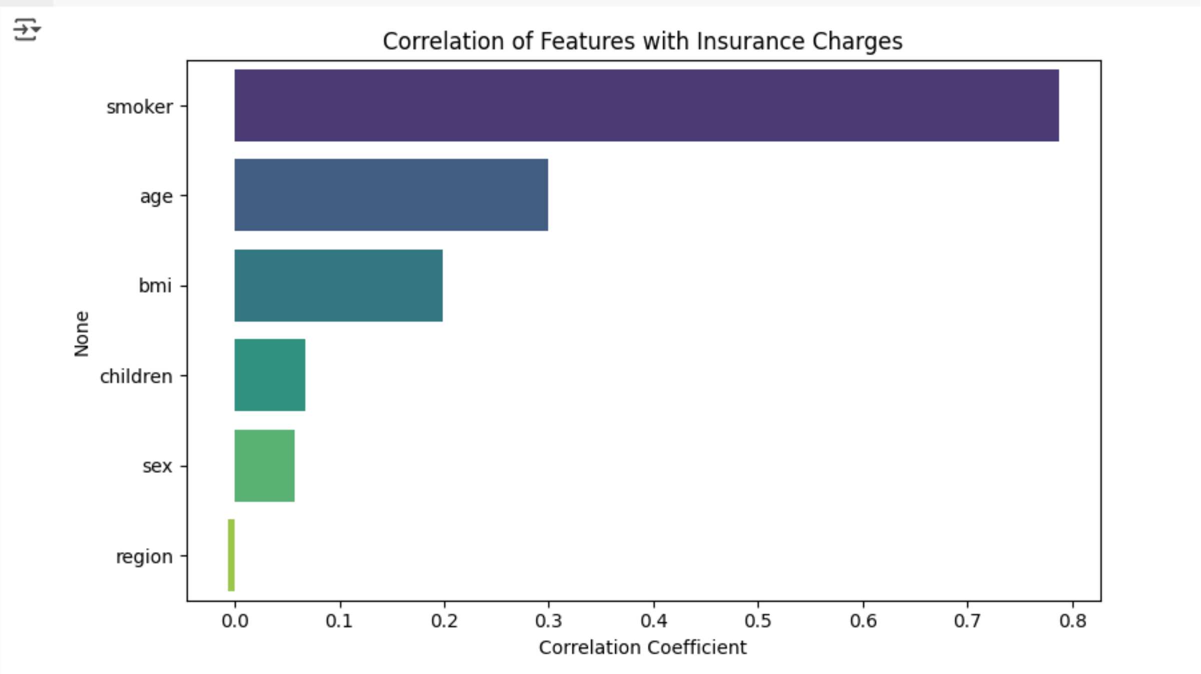
Categorical features such as `sex`, `smoker`, and `region` were encoded using Scikit-learn's `LabelEncoder`. This allowed for the creation of a correlation heatmap and numerical modeling.

Exploratory Analysis

A correlation analysis showed that smoking status had the strongest relationship with insurance charges (r = 0.79), followed by age (r = 0.30) and BMI (r = 0.20). Other variables, such as number of children, sex, and region, showed weak or negligible correlation.



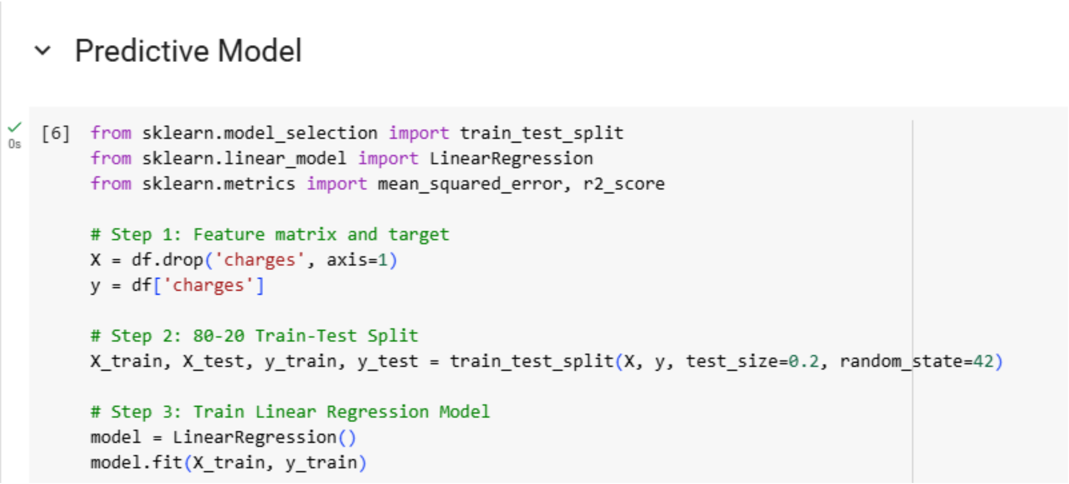
The heatmap clearly shows that **smoking status** has the most substantial influence on insurance charges. This is followed by age and BMI. Other features such as sex and region show minimal correlation. This insight will help us focus on the most relevant predictors when modeling.



The bar chart reinforces the findings from the heatmap: smoking status is by far the most influential predictor of insurance charges, followed by age and BMI. The number of children, sex, and region show very weak or negligible correlations.

Results

Model Training

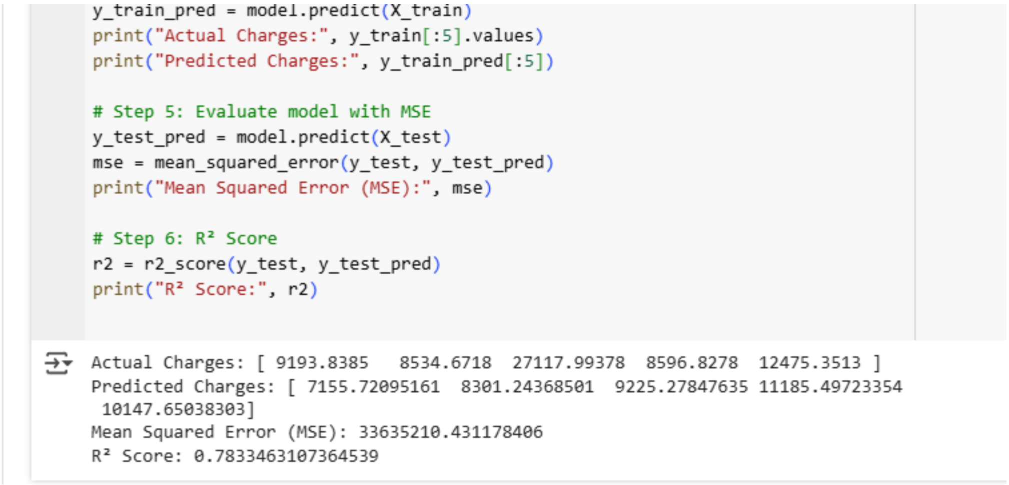


```
from sklearn.model_selection import train_test_split
from sklearn.linear_model import LinearRegression
from sklearn.metrics import mean_squared_error, r2_score

# Step 1: Feature matrix and target
X = df.drop('charges', axis=1)
y = df['charges']

# Step 2: 80-20 Train-Test Split
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2, random_state=42)

# Step 3: Train Linear Regression Model
model = LinearRegression()
model.fit(X_train, y_train)
```

An 80/20 train-test split was used. A linear regression model was trained on the training set to predict insurance `charges`.

Model Performance

- **Mean Squared Error (MSE):** 33,635,210.43
- **R² Score:** 0.7833



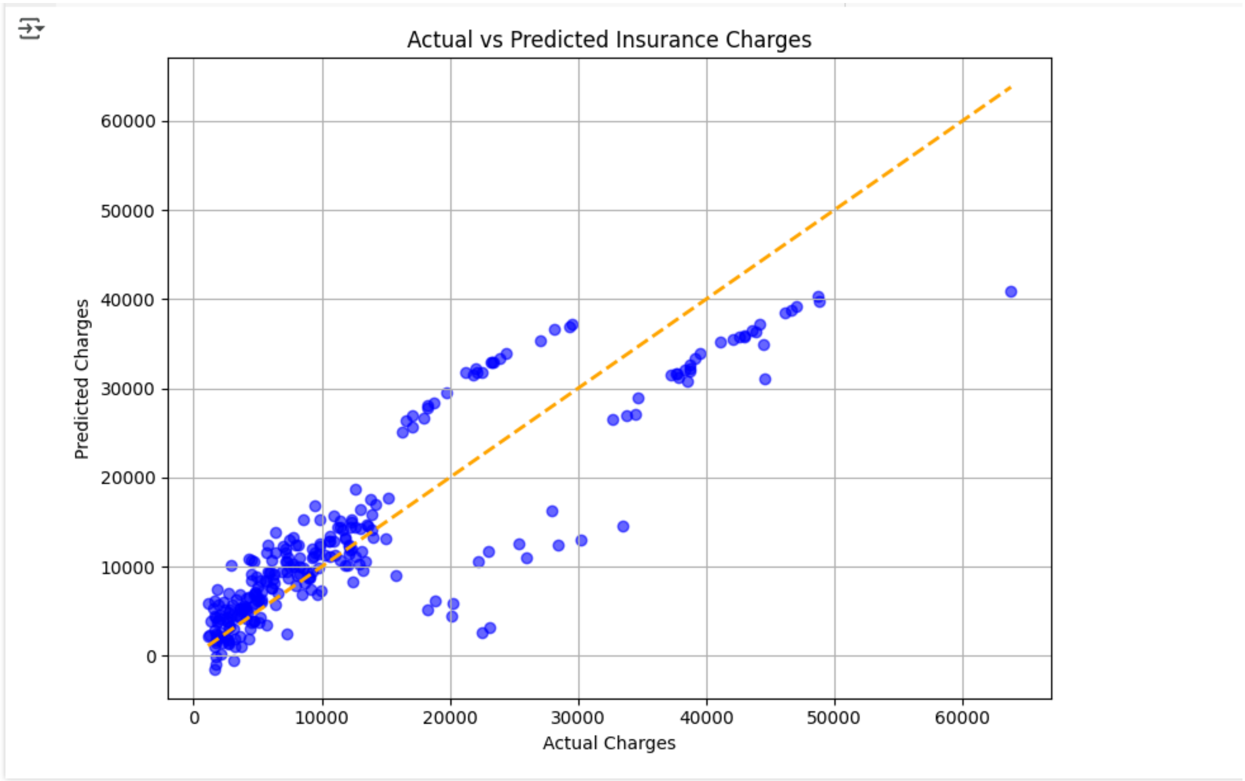
```
y_train_pred = model.predict(X_train)
print("Actual Charges:", y_train[:5].values)
print("Predicted Charges:", y_train_pred[:5])

# Step 5: Evaluate model with MSE
y_test_pred = model.predict(X_test)
mse = mean_squared_error(y_test, y_test_pred)
print("Mean Squared Error (MSE):", mse)

# Step 6: R² Score
r2 = r2_score(y_test, y_test_pred)
print("R² Score:", r2)
```

```
Actual Charges: [ 9193.8385   8534.6718  27117.99378  8596.8278  12475.3513 ]
Predicted Charges: [ 7155.72095161  8301.24368501  9225.27847635 11185.49723354
 10147.65038303]
Mean Squared Error (MSE): 33635210.431178406
R² Score: 0.7833463107364539
```

These results indicate that the model explains approximately 78.3% of the variation in medical charges.

Visualization

The scatter plot below compares actual charges to predicted charges from the model. Points along the orange dashed line represent perfect predictions.



Most points fall near the line, especially in the low to mid charge range**,** showing that the model makes reasonable predictions for typical cases. However, there are a few notable deviations for higher charges, suggesting that the model struggles to fully capture the complexity of more expensive cases. This may be due to non-linear relationships not modeled by simple linear regression or outliers or rare combinations of features (e.g., older smokers with high BMI)

Discussion

The regression model confirms that **smoking status** is the most influential factor affecting medical insurance charges. While the model performs reasonably well, it tends to underpredict very high charges—likely due to non-linear relationships or interactions between features not captured by a simple linear model. The high MSE value further suggests variability that could be better modeled using more advanced techniques.

Conclusion

This project demonstrates the effectiveness of linear regression in modeling insurance charges with a relatively high degree of accuracy. However, to improve predictive performance—especially for outlier cases—future work could involve using polynomial regression, regularization techniques (e.g., Ridge or Lasso), or ensemble methods like Random Forests.

References

- Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., ... & Duchesnay, E. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12, 2825–2830.
- McKinney, W. (2010). Data Structures for Statistical Computing in Python. *Proceedings of the 9th Python in Science Conference*, 51–56.