Health Insurance Cost Prediction Using Machine Learning

Project Report

---

1. Introduction

Healthcare costs have been rising steadily over the past decades, making health insurance an essential aspect of financial planning. Understanding what drives these costs can help both insurers and individuals make better decisions. This project examines various factors that influence health insurance premiums and develops predictive models to estimate costs based on personal attributes.

The dataset used contains information about insurance beneficiaries including their age, gender, BMI, number of dependents, smoking habits, and residential region. Our goal was to identify which factors most significantly impact insurance charges and build accurate prediction models.

---

2. Dataset Overview

The insurance dataset consists of 1,338 records with seven attributes:

- **age**: Age of the primary beneficiary
- **sex**: Gender (male/female)
- **bmi**: Body Mass Index, an indicator of body weight relative to height
- **children**: Number of dependents covered
- **smoker**: Whether the person smokes (yes/no)
- **region**: Geographic location in the US (northeast, southeast, southwest, northwest)
- **charges**: Annual medical costs billed by insurance (target variable)

One positive aspect of this dataset is the complete absence of missing values, which simplified our preprocessing steps and allowed us to focus on analysis and modeling.

---

3. Exploratory Data Analysis

3.1 Distribution of Insurance Charges

The initial examination of insurance charges revealed a right-skewed distribution. Most people have lower insurance costs, but there's a long tail of individuals with very high charges. When we applied a logarithmic

transformation, the distribution became more normalized, which can be beneficial for certain modeling approaches.

3.2 Regional Patterns

Breaking down charges by region showed some interesting patterns:

- The Southeast region has the highest total insurance charges
- The Southwest has the lowest overall costs
- However, these differences need to be interpreted carefully alongside other factors

3.3 Impact of Smoking

Smoking emerged as a dominant factor right from the exploratory phase. When we compared charges between smokers and non-smokers across different regions:

- Smokers consistently face much higher insurance costs
- The Southeast has the highest charges for smokers
- The gap between smokers and non-smokers is substantial across all regions

3.4 Age, BMI, and Insurance Costs

We observed positive correlations between age and charges, as well as BMI and charges. However, these relationships are much stronger for smokers than non-smokers. This suggests an interaction effect where smoking amplifies the impact of other risk factors.

3.5 Family Size Effects

Interestingly, the number of children showed a relatively weak relationship with insurance charges. There's a slight tendency for people with children to have higher costs, but it's not as pronounced as other factors. We also noticed that people with more children tend to smoke less, which could be related to family health considerations.

---

4. Data Preprocessing

Since machine learning algorithms work with numerical data, we needed to convert categorical variables. We used Label Encoding for three categorical features:

- sex (male/female → 0/1)
- smoker (yes/no → 0/1)
- region (four regions → 0/1/2/3)

After encoding, we examined correlations between all variables. The correlation matrix confirmed what we observed visually: smoking has the strongest correlation with charges (around 0.79), followed by age and BMI. Sex, number of children, and region showed weak correlations with the target variable.

---

5. Model Development and Evaluation

We implemented five different regression models to predict insurance charges:

5.1 Linear Regression

As our baseline model, linear regression assumes a straightforward linear relationship between features and charges. The model achieved an R^2 score of approximately 0.75, meaning it explains about 75% of the variance in insurance costs. While this is decent, there's clearly room for improvement.

5.2 Ridge Regression

Ridge regression adds L2 regularization to prevent overfitting. We used an alpha value of 0.5 to control the regularization strength. The performance was very similar to standard linear regression, suggesting that overfitting wasn't a major issue with our relatively small feature set.

5.3 Lasso Regression

Lasso regression uses L1 regularization, which can reduce some coefficients to zero, effectively performing feature selection. With alpha set to 0.2, the model performed comparably to Ridge regression. The fact that all regularized models showed similar performance to linear regression indicates that our features are genuinely predictive rather than contributing noise.

5.4 Random Forest Regressor

This ensemble method builds multiple decision trees and averages their predictions. We trained the model with 100 trees and achieved an R^2 score around 0.86. The improvement over linear models suggests that there are non-linear relationships in the data that tree-based methods can capture better.

The Random Forest model also provided feature importance scores:

1. Smoking status: Most important by far
2. Age: Second most influential
3. BMI: Third in importance
4. Children, sex, and region: Relatively minor contributions

The residual plot for Random Forest showed fairly random scatter around zero, indicating good model fit without obvious patterns in the errors.

5.5 Polynomial Regression

To capture non-linear relationships while staying within the linear regression framework, we created polynomial features up to degree 2. This means we included squared terms and interaction terms between features. We removed sex and region before creating polynomial features since they showed low importance.

This approach worked remarkably well, achieving an R^2 score around 0.87, slightly better than Random Forest. The model's predictions closely tracked actual values, with most points falling near the perfect prediction line.

Model evaluation metrics for Polynomial Regression:

- Mean Absolute Error: Around $2,500-3,000
- Root Mean Squared Error: Around $4,500-5,000

These numbers mean that on average, our predictions are within about $2,500-3,000 of the actual cost, which is reasonable given the wide range of insurance charges in the dataset.

---

6. Results Summary

Comparing all models:

Model	R^2 Score
Polynomial Regression	0.870
Random Forest	0.860
Linear Regression	0.751
Ridge Regression	0.750
Lasso Regression	0.749

Polynomial Regression emerged as the best performer, though Random Forest was very close. The simpler linear models, while less accurate, still captured the major patterns in the data.

---

7. Key Findings

Based on our analysis, we can draw several conclusions:

Primary Finding: Smoking is overwhelmingly the most important factor in determining health insurance costs. Smokers pay substantially more than non-smokers, and this effect compounds with other risk factors like age and BMI.

Secondary Factors: Age and BMI also contribute significantly to costs. As people get older or have higher BMI values, their insurance charges tend to increase. However, these effects are much more pronounced for

smokers.

Minimal Impact: Gender, number of children, and geographic region have relatively small effects on insurance costs when other factors are controlled for.

Model Performance: Non-linear models (Polynomial Regression and Random Forest) outperformed linear models, suggesting that the relationships between features and insurance costs are more complex than simple linear associations.

---

8. Practical Implications

These findings have real-world applications:

For Insurance Companies: The models can help set more accurate premiums based on individual risk factors. The strong influence of smoking justifies higher premiums for smokers.

For Individuals: Understanding these factors can motivate lifestyle changes. Quitting smoking would have the largest impact on reducing insurance costs, followed by maintaining a healthy weight.

For Policy Makers: The data highlights the substantial healthcare burden associated with smoking, which could inform public health initiatives and tobacco control policies.

---

9. Limitations and Future Work

This analysis has some limitations worth noting:

- The dataset is relatively small (1,338 records), which may not capture all the diversity in the population
- We don't have information about pre-existing conditions, which likely affect insurance costs
- The temporal aspect is missing - we don't know how costs change over time for the same individuals
- Geographic information is limited to broad regions rather than specific states or cities

Future improvements could include:

- Gathering more data to improve model robustness
- Adding features like medical history and occupation
- Exploring deep learning approaches if more data becomes available
- Developing separate models for different subgroups (e.g., smokers vs. non-smokers)
- Creating a web interface for easy predictions

---

10. Conclusion

This project successfully developed predictive models for health insurance costs with reasonable accuracy. The Polynomial Regression model achieved the best performance with an R^2 score of 0.87, meaning it can explain 87% of the variation in insurance charges.

The analysis clearly demonstrated that smoking is the dominant factor affecting health insurance costs, followed by age and BMI. These findings align with medical knowledge about health risks and validate our modeling approach.

The models developed here could be useful tools for insurance companies to estimate premiums and for individuals to understand how different factors affect their insurance costs. While there's always room for improvement, the current models provide a solid foundation for insurance cost prediction.

References

- Scikit-learn Documentation: Machine Learning in Python
- Seaborn: Statistical Data Visualization
- Pandas: Data Analysis Library
- NumPy: Numerical Computing Tools

Project completed as part of Machine Learning coursework