Project Report:

Medical Health Expense Prediction

GROUP - 11

TEAMMATES :

KAZI SAJID MAHMUD (2013388042)

SHADMAN SHARIAR (1911457642)

ANIK PAL (2012673042)

Table of Contents

1. Introduction

The rising cost of healthcare is a significant concern for both individuals and insurance companies. So we make a accurate prediction of medical expenses can aid in better financial planning and risk management. This project aims to develop a predictive model to estimate individual medical expenses using various demographic and lifestyle factors. Understanding these expenses is crucial for insurance companies to set premiums and for individuals to anticipate their healthcare costs.

2. Data Description

The dataset used in this project includes several features that are likely to influence medical expenses:

- **Age**: Age of the primary beneficiary.
- **Sex**: Gender of the beneficiary (male/female).
- **BMI**: Body Mass Index, a measure of body fat based on height and weight.
- **Children**: Number of children/dependents covered by the insurance plan.
- **Smoker**: Smoking status of the beneficiary (smoker/non-smoker).
- **Region**: Geographic region of the beneficiary (northeast, northwest, southeast, southwest).
- **Charges**: Annual medical expenses billed by health insurance.

The dataset is comprehensive and includes both numerical and categorical variables, which necessitates appropriate preprocessing steps before model training.

3. Exploratory Data Analysis (EDA)

Exploratory Data Analysis is essential to understand the underlying patterns and relationships in the data.

Distribution of Medical Charges

The distribution of medical charges is right-skewed, indicating that most individuals incur lower costs, while a few incur very high expenses. This skewness is a critical observation as it can impact the model's performance and may require transformation for better predictions.

Relationship Between Features and Medical Charges

Age

Older individuals tend to have higher medical expenses. This trend is expected as the likelihood of health issues increases with age, leading to higher medical costs.

BMI

Higher BMI values are associated with increased medical costs, particularly for smokers. This relationship highlights the impact of obesity and smoking on health, leading to higher insurance claims.

Number of Children

The number of children has a minimal impact on medical charges. This observation might be surprising initially, but it suggests that the primary factor driving medical costs is the health status of the policyholder rather than the number of dependents.

Smoking Status

Smokers incur significantly higher medical costs compared to non-smokers. This stark difference underscores the health risks associated with smoking and its financial implications on medical expenses.

Region

The region has a slight impact on medical expenses, with some regions showing higher average costs. Regional differences might be due to varying healthcare costs and availability of medical services across different areas.

4. Data Preprocessing

Data preprocessing involves preparing the data for modeling by handling missing values, encoding categorical variables, and scaling numerical features.

Handling Missing Values

The dataset does not contain any missing values, which simplifies the preprocessing steps.

Encoding Categorical Variables

Categorical variables such as sex, smoker, and region are encoded using one-hot encoding to convert them into numerical format. This step is crucial for the model to process categorical data appropriately.

Feature Scaling

Numerical features are scaled to ensure they contribute equally to the prediction model. Feature scaling helps in improving the convergence rate of gradient descent-based algorithms and ensures that no particular feature dominates the others due to its scale.

5. Model Development

A linear regression model is developed to predict medical expenses based on the provided features. Linear regression is chosen for its simplicity and interpretability, providing a baseline model for this prediction task.

Splitting the Data

The data is split into training and testing sets to evaluate the model's performance. An 80-20 split is typically used, where 80% of the data is used for training and 20% for testing.

Training the Model

A linear regression model is trained using the training data. The training process involves fitting the model to minimize the difference between the predicted and actual medical charges.

6.Transition to Online Learning

To enhance the model's adaptability and efficiency, we converted our machine learning project from batch learning to online learning. Online learning processes data incrementally, allowing the model to update continuously as new data becomes available. This approach is particularly beneficial for real-time applications and environments where data is continuously generated.

7. Model Evaluation

The model is evaluated using several metrics to assess its performance.

Evaluation Metrics

The model's performance is assessed using Mean Absolute Error (MAE), Mean Squared Error (MSE), and R-squared. These metrics provide a comprehensive view of the model's accuracy and its ability to explain the variance in the data.

Mean Absolute Error (MAE)

MAE measures the average absolute difference between the predicted and actual medical charges, providing an intuitive measure of prediction accuracy.

Mean Squared Error (MSE)

MSE measures the average squared difference between the predicted and actual medical charges, penalizing larger errors more than smaller ones. It is useful for understanding the overall prediction error.

R-squared

R-squared indicates the proportion of variance in the dependent variable that is predictable from the independent variables. A higher R-squared value indicates a better fit of the model to the data.

8. Results

The results of the model evaluation indicate how well the model predicts medical expenses.

- **Mean Absolute Error (MAE)**: The model's MAE provides an average measure of how much the predicted charges deviate from the actual charges.
- **Mean Squared Error (MSE)**: The MSE highlights the overall error magnitude, with larger errors being more heavily penalized.
- **R-squared**: The R-squared value provides insight into the proportion of variance in medical charges explained by the model.

The linear regression model, while basic, offers a foundational approach to understanding the relationships between the features and medical expenses. However, the performance metrics suggest that there is room for improvement.

9. Conclusion

The linear regression model provides a basic approach to predicting medical expenses. While the model shows some predictive capability, there is potential for improvement. The model's performance metrics suggest that more sophisticated algorithms and additional feature engineering could enhance its accuracy.

The findings highlight the significant impact of factors such as age, BMI, and smoking status on medical expenses. Understanding these relationships can help in better managing healthcare costs and setting insurance premiums.

10. Future Work

To improve the predictive accuracy of the model, several steps can be taken:

- **Implementing Advanced Machine Learning Algorithms**: Using algorithms such as decision trees, random forests, gradient boosting machines, or neural networks.
- **Hyperparameter Tuning**: Fine-tuning the model parameters to achieve optimal performance.
- **Feature Engineering**: Exploring additional features that may impact medical expenses, such as lifestyle factors, genetic predispositions, and pre-existing health conditions.
- **Cross-validation**: Using cross-validation techniques to ensure the model's robustness and generalizability.
- **Ensemble Methods**: Combining multiple models to improve prediction accuracy.

The project demonstrates the potential of predictive modeling in the healthcare domain, providing a foundation for further exploration and refinement in predicting medical expenses. Future work will focus on enhancing the model's accuracy and exploring additional factors that influence medical costs.

11. References

1. Y. Bengio, A. Courville, and P. Vincent, "Representation Learning: A Review and New Perspectives," in IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 35, no. 8, pp. 1798-1828, Aug. 2013, doi: 10.1109/TPAMI.2013.50.
2. A. Rajkomar, J. Dean, and I. Kohane, "Machine Learning in Medicine," in New England Journal of Medicine, vol. 380, pp. 1347-1358, 2019, doi: 10.1056/NEJMra1814259.
3. G. E. Hinton and R. R. Salakhutdinov, "Reducing the Dimensionality of Data with Neural Networks," in Science, vol. 313, no. 5786, pp. 504-507, July 2006, doi: 10.1126/science.1127647.
4. P. B. Jensen, L. J. Jensen, and S. Brunak, "Mining Electronic Health Records: Towards Better Research Applications and Clinical Care," in Nature Reviews Genetics, vol. 13, pp. 395-405, 2012, doi: 10.1038/nrg3208.
5. K. P. Murphy, "Machine Learning: A Probabilistic Perspective," MIT Press, 2012. (ISBN: 978-0-262-01802-9)
6. D. Bertsimas, V. R. Gupta, and J. Kallus, "Data-Driven Robust Optimization," in Mathematical Programming, vol. 167, no. 2, pp. 235-292, 2018, doi: 10.1007/s10107-017-1124-8.
7. J. C. Platt, "Sequential Minimal Optimization: A Fast Algorithm for Training Support Vector Machines," Microsoft Research, 1998.
8. I. Guyon and A. Elisseeff, "An Introduction to Variable and Feature Selection," in Journal of Machine Learning Research, vol. 3, pp. 1157-1182, 2003.
9. T. Chen and C. Guestrin, "XGBoost: A Scalable Tree Boosting System," in Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 785-794, 2016, doi: 10.1145/2939672.2939785.
10. A. Hyafil and R. L. Rivest, "Constructing Optimal Binary Decision Trees is NP-Complete," in Information Processing Letters, vol. 5, no. 1, pp.

4o