Medical Insurance Cost Prediction Using Machine Learning

Goal of the Project

The goal of this project is to create an automated system that can predict individual medical insurance costs using machine learning. This predictive model enables insurance companies to estimate the likely cost for prospective clients based on their personal and health characteristics.

Problem Statement

A medical insurance company needs a machine learning model that can predict the medical insurance cost for an individual based on their personal demographics and lifestyle factors. This system should automate the insurance cost estimation process to enhance accuracy and efficiency in pricing policies.

Dataset Overview

The dataset used in this project contains information on health and demographic attributes that impact insurance costs. Originating from *Machine Learning with R* by Brett Lantz, this dataset includes seven features, with the "charges" column as the target variable to be predicted.

Features

1. **age**: Age of the primary beneficiary, which can significantly affect medical costs.
2. **sex**: Gender of the insurance policyholder, represented as male or female.
3. **bmi**: Body Mass Index, calculated as kg/m², indicating if the individual's weight is healthy relative to their height.
4. **children**: Number of dependents covered under the insurance plan.
5. **smoker**: Indicates if the individual smokes, which typically increases health risks and insurance costs.
6. **region**: Residential area in the United States, categorized as northeast, southeast, southwest, or northwest.
7. **charges**: The medical costs billed by health insurance, which is the target variable for prediction.

Workflow

1. **Data Collection**: The dataset was sourced from Kaggle, containing features that influence health costs.
2. **Data Analysis**: Exploratory data analysis (EDA) was conducted to understand the data distribution, identify patterns, and evaluate relationships between features and the target variable.
3. **Data Preprocessing**: Since the data includes categorical features (e.g., sex, smoker, region), encoding was required to convert these to numeric values. This step ensures the data is in a format suitable for machine learning models.
4. **Data Split**: The dataset was divided into training and testing sets, with an 80-20 ratio to allow for effective model evaluation.
5. **Model Training with Linear Regression**: Linear Regression was initially chosen as a baseline model to assess its ability to capture relationships within the dataset.
6. **Model Testing and Evaluation**: The trained model was evaluated on the test set to check its prediction accuracy.
7. **Advanced Models with XGBoost and CatBoost**: Due to the limitations observed with Linear Regression, XGBoost and CatBoost models were applied to improve accuracy.
8. **Model Deployment**: The final model was deployed on Hugging Face using Gradio to create a user-friendly interface for real-time predictions.

Machine Learning Models

Linear Regression Model

Why Linear Regression?

Linear Regression was chosen as the initial model because of its simplicity and effectiveness with datasets exhibiting linear relationships. It served as a baseline to understand how well a basic model could predict insurance costs.

Performance

While Linear Regression provided reasonable predictions, the R-squared value was lower compared to more complex models. Linear Regression struggled with capturing non-linear patterns and feature interactions, limiting its accuracy for this dataset.

XGBoost Model

Why XGBoost?

To improve prediction performance, XGBoost was applied as it handles non-linear relationships and interactions between features better than Linear Regression. XGBoost's ability to model complex patterns is ideal for datasets with a mix of categorical and continuous variables.

Performance

XGBoost achieved high accuracy on the training data with an almost perfect R-squared score, but the test set accuracy indicated overfitting. The model performed well on training data but didn’t generalize as effectively on unseen data, likely due to high complexity.

CatBoost Model

Why CatBoost?

CatBoost was selected as it specializes in handling categorical data, an advantage given the presence of sex, smoker, and region columns. CatBoost requires minimal preprocessing of categorical features, making it suitable for this dataset.

Performance

CatBoost demonstrated a balanced performance on both training and test data, with high R-squared scores and minimal overfitting. It achieved better generalization compared to XGBoost, making it the final model choice for deployment.

Model Deployment with Gradio

Gradio Interface

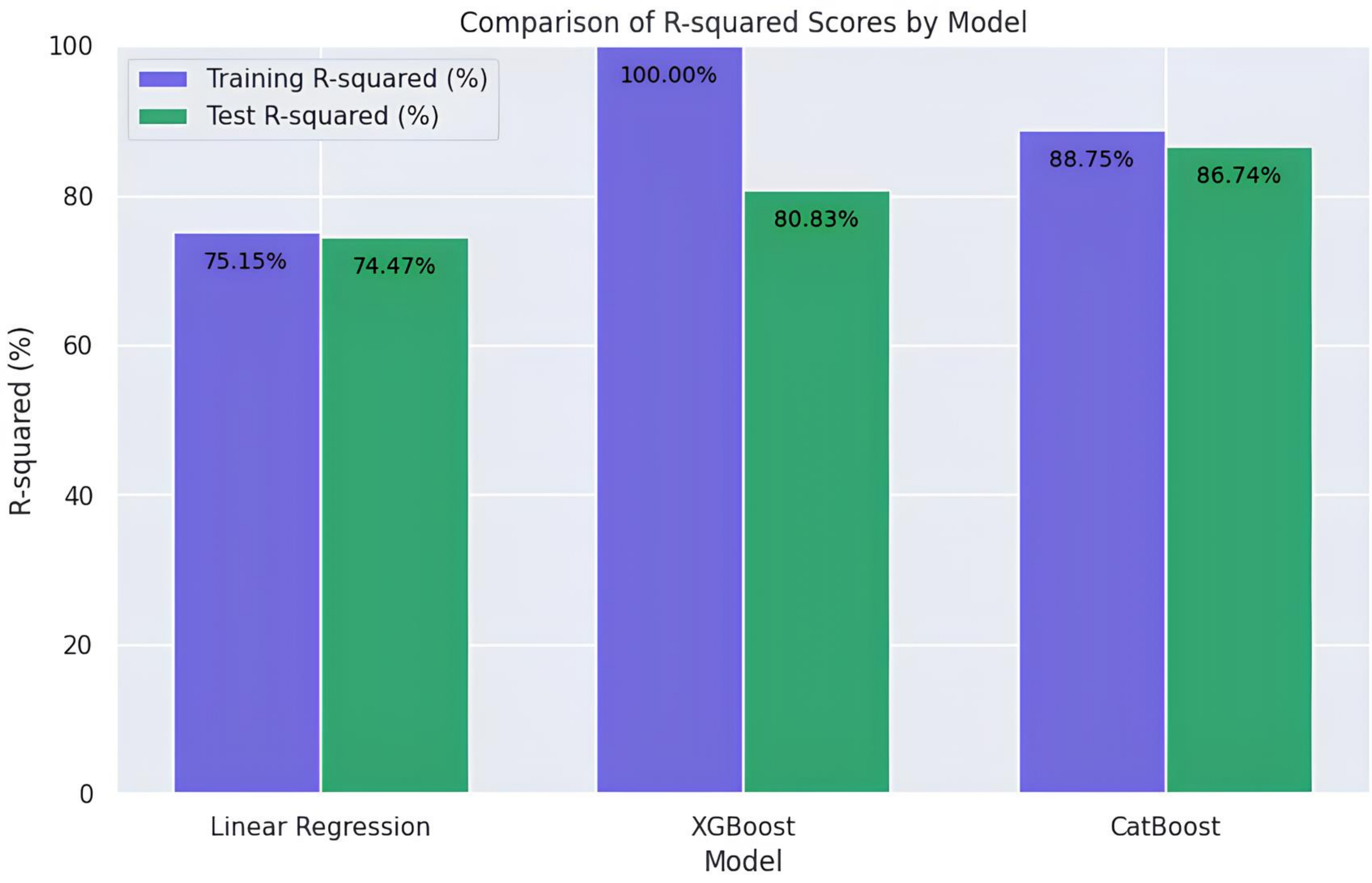
Gradio was used to build an interactive interface, allowing users to input features (age, sex, BMI, children, smoker status, region) and get immediate insurance cost predictions. This user-friendly interface enables broader accessibility, allowing non-technical users to interact with the model seamlessly.

Hosting on Hugging Face

The Gradio app was hosted on Hugging Face, making the model accessible to a wider audience without the need for complex infrastructure. Hosting on Hugging Face ensures scalability and easy access to the predictive system.

Conclusion

This project demonstrates the process of building, evaluating, and deploying a machine learning model for medical insurance cost prediction. Starting with Linear Regression as a baseline, the project explored more complex algorithms like XGBoost and CatBoost to achieve higher accuracy. CatBoost was chosen for deployment due to its strong generalization capabilities. The final Gradio app, hosted on Hugging Face, provides an accessible solution for insurance cost estimation based on user input.



Data Set Link:

https://www.kaggle.com/datasets/mirichoi0218/insurance

Gradio App Link:

https://huggingface.co/spaces/Sourudra/Medical_Insurance_Cost_Predictor

GitHub Link:

https://github.com/Sourudra/Medical-Insurance-Cost-Prediction-With-Machine-Learning