INNOMATİCS
RESEARCH LABS

Machine Learning - Project Report Document

Student Name	Sada Vijay
Batch	AI Elite 18
Project Name	Medical Cost Prediction
Project Domain	Healthcare
Type of Machine Learning	Supervised ML
Type of Problem	Regression
Project Methodology	CRISP-DM
Stages Involved	• Data Collection and Understanding • Data Preparation • Model Building • Model Training • Model Evaluation

Business Understanding:

The concept of insurance involves a contractual agreement between two parties, where one assumes the risk of the other in exchange for a fee called a premium, and pledges to compensate the insured party upon the occurrence of an uncertain event. Specifically focusing on health insurance, it refers to a scheme that either covers or contributes towards the expenses associated with medical care.

Health insurance is a growing market in India, driven by factors like increasing middle class, rising healthcare costs, and growing awareness. This suggests a significant opportunity for companies offering health insurance products.

Insurance companies need to predict the average medical costs of future insured individuals by analyzing population trends, even though they have limited data on each individual. This is crucial to setting profitable insurance premiums.

Problem Statement: This project aims to develop a model that can predict the medical charges an individual will incur, based on available data, allowing for more accurate premium pricing.

Here are some potential business constraints:

1. Regulatory Compliance
2. Data Privacy and Security
3. Resource Limitations
4. Accuracy and Reliability
5. Interpretability
6. Ethical Considerations
7. Competitive Landscape

Accuracy vs. Interpretability: Highly accurate models can be complex and challenging to explain, hindering understanding of cost variations by customers or regulators.

External factors: The model may not consider sudden changes in healthcare regulations, drug prices, or new disease outbreaks.

Regulatory Compliance: Adhering to insurance industry regulations and privacy laws may limit data usage and model deployment.

Data Privacy and Security: Ensuring customer data privacy and security is essential, potentially impacting data handling and model development due to compliance requirements.

Stage 1: Data Collection and Understanding

a) **Data Collection:** The data has been provided to us by the client.

b) **Data Understanding:**

1. Age: The age of the primary beneficiary.
2. Sex: The gender of the insurance contractor, categorized as female or male.
3. BMI: Body Mass Index, indicating the relative body weight based on height, an objective measure of body weight in kilograms per square meter (kg/m^2). Ideally falls within the range of 18.5 to 24.9.
4. Children: The number of dependents covered by the insurance.
5. Smoker: Indicates whether the individual smokes or not.
6. Region: The residential area of the beneficiary within the United States, categorized into northeast, southeast, southwest, or northwest.
7. Charges: The individual medical costs billed by the health insurance provider.

S No	Feature Name	Data Type
1	age	int64
2	sex	Object
3	bmi	float64
4	children	int64
5	smoker	Object
6	region	Object
7	charges	float64

Stage 2: Data Preparation

a) Exploratory Data Analysis:

S No	Type	Feature Names	Observation
1	Missing Values	NA	NA
2	Duplicates	Row index: 581	There exist a duplicate datapoint.
3	Outliers	bmi	There exist outliers in this column.

b) Data Cleaning/wrangling:

S no	Type of Cleaning	Technique	Feature Name	Reason
1	Duplicate value	Drop	Row index: 581	To maintain data consistency and accuracy
2	Encoding	One hot	Sex, smoker, region	Used One Hot Encoding since the data in these categorical columns are nominal.
3	Scaling	Robust Scaling	Age, bmi, children	Used Robust Scaling since there exists outliers in the column “bmi” and robust scaler handles the outliers better.

Stage 3: Model Building:

S No	Type of Problem	Algorithm Name
1	Regression	Linear Regression
2	Regression	KNeighbors Regressor
3	Regression	Decision Tree Regressor
4	Regression	Random Forest Regressor
5	Regression	SVR

1. **Linear Regression:** Linear regression is a statistical method used to model the relationship between a dependent variable and one or more independent variables by fitting a linear equation to observed data points. It's commonly employed for predicting continuous outcomes based on linear relationships between variables.

2. **KNeighbors Regressor:** K-nearest neighbors (KNN) regression is a non-parametric regression technique used for estimating the value of a target variable by averaging the values of its k-nearest neighbors. It relies on the assumption that similar data points have similar target values and is particularly effective for locally smooth relationships between features and the target variable.

3. **Decision Tree Regressor:** Decision tree regression is a non-linear regression technique that recursively splits the data into subsets based on the values of input features, aiming to predict the target variable within each subset. It constructs a tree-like structure where each internal node represents a decision based on a feature, leading to a final prediction at the leaf nodes.

4. **Random Forest Regressor:** Random Forest Regressor is an ensemble learning technique that combines multiple decision trees to improve predictive accuracy and reduce overfitting in regression tasks. It aggregates predictions from individual trees and outputs the average prediction, making it robust and effective for modeling complex relationships in data.

5. **SVR:** Support Vector Regression (SVR) is a regression algorithm that utilizes support vector machines to find a hyperplane that best fits the data points while maximizing the margin. It is particularly effective in capturing non-linear relationships and handling high-dimensional data, making it suitable for various regression tasks.

Stage 4: Model Training:

S No	Algorithm Name	Metric used for Evaluation
1	Linear Regression	Mean Absolute Error
2	KNeighbors Regressor	Mean Absolute Error
3	Decision Tree Regressor	Mean Absolute Error
4	Random Forest Regressor	Mean Absolute Error
5	SVR	Mean Absolute Error

Stage 5: Model Evaluation:

S No	Algorithm Name	Metric Score
1	Linear Regression	4323.963298
2	KNeighbors Regressor	3854.502497
3	Decision Tree Regressor	3081.074484
4	Random Forest Regressor	2748.811625
5	SVR	9114.819321

Challenges Faced:

We've found that there are outliers in the input column "bmi". As they are true outliers we've retained the values and as we can't use standard scaler for this data since it might create a bias due to the outliers, so in place of it we've used the robust scaler since it handles the outliers better.

Conclusion:

We can observe that out of all the models that we've trained, the Random Forest Regressor model is giving the lowest mean absolute error, which means predictions made by this model are the closest to the real values in comparison to the other models. So, we can say that by observing the evaluation metric of all models that the Random Forest Regressor is the best algorithm for the Medical Cost Prediction problem.