**Machine Learning & Algorithms -1**

**CIA-1**

# **Application of Various Regression Models in Insurance Sector**

**By**

**Ankit Ray-2327510**

**Under the Guidance of**

**Dr.HELEN JOSEPHINE V L**

****

**MBA PROGRAMME**

**SCHOOL OF BUSINESS AND MANAGEMENT**

**CHRIST (DEEMED TO BE UNIVERSITY), BANGALORE**

**July 2024**

# 

### 

### **1. Business Understanding**

#### **a. Problem Identification**

* Problem: Forecasting insurance charges based on demographic and lifestyle factors.

#### **b. Variables**

* Predictors (Independent Variables):
  + age: The age of the insured person, which may correlate with health risks and medical costs.
  + sex: Gender of the insured, which could affect health risk profiles.
  + bmi: Body Mass Index (BMI), a measure of body fat based on height and weight. Higher BMI may indicate higher health risks.
  + children: Number of children/dependents covered under the insurance, affecting family healthcare costs.
  + smoker: Binary variable indicating if the insured is a smoker, which significantly impacts health risks and insurance costs.
  + region: Geographic region of the insured, influencing healthcare costs due to regional variations in medical expenses and regulations.
* Target (Dependent Variable):
  + charges: Insurance charges or premiums paid by the insured, influenced by the above predictors and other factors.

#### **c. Objectives**

* Develop a predictive model using Multiple Linear Regression to accurately estimate insurance charges based on demographic and lifestyle data.
* Provide insights into how different factors influence insurance charges, assisting in risk assessment and pricing strategies.

### **2. Data Understanding**

#### **a. Data Collection**

* Describe the process of collecting the insurance dataset containing demographic and charge information, ensuring data privacy and regulatory compliance.

#### **b. Data Exploration**

* Explore the dataset to understand its structure, distributions, and relationships among variables using statistical summaries and visualizations (e.g., histograms, scatter plots).

#### **c. Assessing Data Quality**

* Evaluate data quality by checking for completeness (missing values), correctness (data accuracy), and consistency (uniform data formats and scales).

### **3. Data Preparation**

#### **a. Data Integration**

* Confirm if the dataset required integration from multiple sources (if applicable), ensuring data consistency and compatibility.

#### **b. Data Cleaning**

##### **i. Missing Value Analysis**

* Identify and handle missing values in the dataset using appropriate techniques like imputation (e.g., mean, median).

##### **ii. Data Imputation**

* Impute missing values for variables such as BMI or children using statistical methods to preserve data integrity.

##### **iii. Variable Standardization**

* Standardize numerical variables like age and BMI to a common scale (e.g., z-scores) to facilitate comparison and model convergence.

##### **iv. Feature Selection/Engineering**

* Perform feature engineering if needed, creating new variables such as BMI categories (e.g., underweight, normal, overweight) to capture nonlinear relationships.

##### **v. Outliers Detection and Treatment**

* Detect outliers in insurance charges or other variables using statistical methods (e.g., box plots, z-scores) and apply appropriate treatments (e.g., Winsorization, transformation) to mitigate their impact on model performance.

### **4. Modeling (Multiple Linear Regression, Lasso Regression, Ridge Regression)**

#### **a. Model Selection and Assumptions**

* Explain why Multiple Linear Regression, Lasso Regression, and Ridge Regression were considered and outline their underlying assumptions (e.g., linearity, independence of errors).

#### **b. Model Output**

* Explanation of the Model Equation: Interpret the Multiple Linear Regression equation predicting charges based on predictors (e.g., charges = β0 + β1 \* age + β2 \* bmi + ... + ε).
* Explanation of Parameters and Coefficients: Interpret the coefficients (β) of each predictor variable, indicating the direction and strength of their impact on insurance charges.
* Model Fit Indices: Include R-squared and possibly adjusted R-squared to assess how well the model fits the data, explaining the variance explained by the predictors.

#### **c. Model Interpretation from a Business Point of View**

* Interpret how each predictor (e.g., age, smoking status) influences insurance charges, providing actionable insights for insurance pricing and risk assessment.

### **5. Model Evaluation and Diagnostics**

* Assess the performance of each model using metrics like R-squared (how well the model explains variance in charges) and Mean Squared Error (MSE, how close predictions are to actual charges).

#### **Comparison of Model Performance**

| **Model** | **R-Squared** | **MSE** | **Recommendation** |
| --- | --- | --- | --- |
| Multiple Linear Regression | 0.750913 | 36,501,893 | Recommended Model |
| Lasso Regression | 0.7506324 | 36,552,437 |  |
| Ridge Regression | 0.7508171 | 37,104,117 |  |

#### Compare the R-squared and MSE values of Multiple Linear Regression, Lasso Regression, and Ridge Regression models to evaluate their performance in predicting insurance charges.

* **Recommended Model**: Multiple Linear Regression is recommended for predicting insurance charges based on its highest R-squared and lowest MSE among the models evaluated.

### **Conclusion**

* Summarize findings from the model evaluation and diagnostics, emphasizing the recommended Multiple Linear Regression model for predicting insurance charges based on its performance metrics and interpretative insights.
* Highlight the implications of the recommended model for stakeholders, supporting informed decision-making in insurance pricing and risk management.

### **6. A Short Note on Democratizing the Solution**

* Describe strategies to make the recommended model accessible and actionable for stakeholders, ensuring transparency and usability in decision support systems.
* Discuss potential challenges and considerations in deploying the model in a business context, addressing regulatory compliance and data privacy concerns.