Insurance Claim Analysis

S Inesh Rao
dept. of Computer Science Engineering
Dayananda Sagar University
Bangalore, Karnaaka, India
inesh.cse20@gmail.com

G Jahnavi
dept. of Computer Science Engineering
Dayananda Sagar University
Bangalore, Karnaaka, India
akulajahnavi08@gmail.com

Abstract— **The Insurance Claim Analysis project aimed to develop a predictive model for estimating insurance claim amounts based on a dataset containing various features such as age, gender, smoker status, and medical history. The analysis involved comprehensive data preprocessing, exploratory data analysis, feature engineering, and model training using Support Vector Machine (SVM), Random Forest Regression, and Linear Regression algorithms. The models were evaluated using important metrics including R2 score, RMSE score, and MAE score. Results demonstrated that the Random Forest Regression model exhibited the highest performance, delivering accurate predictions of insurance claim amounts. This project contributes significantly to the insurance industry by enabling more precise estimation of claim amounts, empowering insurance companies to make informed decisions, optimize their processes, and enhance customer satisfaction.**

Keywords— Insurance, Analysis, SVM, EDA, Random Forest Regression, Linear Regression

I. Introduction

The insurance industry plays a vital role in safeguarding individuals and businesses against financial risks and uncertainties. Insurance companies face the constant challenge of accurately estimating claim amounts, as it directly impacts their operations, financial planning, and customer satisfaction. A precise estimation of claim amounts allows insurance companies to make informed decisions, allocate resources efficiently, and maintain a sustainable business model.

The Insurance Claim Analysis project focuses on developing a predictive model that can accurately estimate insurance claim amounts. By leveraging data analytics and machine learning techniques, this project aims to provide insurance companies with a reliable tool to enhance their claim estimation process. The predictive model takes into account various factors such as age, gender, smoker status, and medical history, which can significantly influence the claim amounts.

The project involves several key steps. Firstly, a comprehensive understanding of the dataset is obtained through exploratory data analysis, including data visualization and statistical summaries. This step helps identify patterns, relationships, and potential outliers within the data. Subsequently, data preprocessing techniques are applied to clean the dataset, handle missing values, and ensure the data is in a suitable format for analysis.

Feature engineering is another crucial aspect of the project, where additional meaningful features are derived or transformed from the existing dataset. This process aims to capture hidden patterns or improve the representation of the data, enhancing the predictive power of the model.

The predictive modeling phase involves training and evaluating different machine learning algorithms, such as Support Vector Machine (SVM), Random Forest Regression, and Linear Regression. These algorithms are capable of learning from historical data to make accurate predictions of insurance claim amounts based on the provided features.

The evaluation of the models is performed using well-established metrics such as R2 score, Root Mean Squared Error (RMSE), and Mean Absolute Error (MAE). These metrics assess the performance of the models in terms of their ability to capture the variance in claim amounts and minimize prediction errors.

By developing a reliable predictive model, insurance companies can enhance their claim estimation process, streamline their operations, and improve customer satisfaction. Accurate claim estimations empower insurance companies to allocate resources effectively, optimize their underwriting and pricing strategies, and provide fair and competitive insurance services to their customers.

II. Literature survey

[2] According to M. A. Hearst, S. T. Dumais, E. Osuna, J. Platt and B. Scholkopf, Support vector machines algorithm are considered as lying at the intersection of learning theory and practice: for certain simple types of algorithms, statistical learning theory can identify rather precisely the factors that need to be taken intoaccount to learn successfully

[3] According to J. K. Jaiswal and R. Samikannu, RF (Random Forest) has emerged as a robust algorithm that can handle a feature selection problem with a higher number of variables. It is also very much efficient while dealing with regression problems.

[4] According to M. Huang, Linear regression refers to the mathematical technique of fitting given data to a function of a certain type. It is best known for fitting straight lines.

III. Dataset Description

The dataset used in this study was obtained from Kaggle [1] and consists of 1340 records representing insurance claims. Each record contains the following attributes:

- 'index': This attribute serves as a unique identifier for each record in the dataset.

- 'PatientID': This attribute represents a unique identifier assigned to each patient.
- 'age': The 'age' attribute indicates the age of the patient in years.
- 'gender': The 'gender' attribute specifies the gender of the patient, with values 'male' or 'female'.
- 'bmi': The 'bmi' attribute represents the Body Mass Index (BMI) of the patient, which is a measure of body fat calculated based on the patient's height and weight.
- 'bloodpressure': The 'bloodpressure' attribute captures the blood pressure reading of the patient.
- 'diabetic': The 'diabetic' attribute indicates whether the patient has been diagnosed with diabetes or not, with values 'yes' or 'no'.
- 'children': The 'children' attribute denotes the number of children the patient has.
- 'smoker': The 'smoker' attribute identifies whether the patient is a smoker or a non-smoker, with values 'yes' or 'no'.
- 'region': The 'region' attribute represents the geographic region where the patient resides.

The target variable of interest in this dataset is 'claim', which denotes the insurance claim amount associated with each record.

This dataset with 1340 records provides valuable information for analysing and modelling insurance claim data, allowing for the development of predictive models to estimate claim amounts accurately.

IV. Data Preprocessing

Data preprocessing is a crucial step in preparing the dataset for analysis. It involves cleaning and transforming the raw data to ensure its quality and suitability for modelling. Tasks such as handling missing values, removing outliers, normalizing or scaling features, and encoding categorical variables are performed to enhance the accuracy and effectiveness of subsequent data analysis and modelling processes.

A. *Loading the Dataset*

The dataset, in CSV format, is loaded into a Pandas Data Frame, providing a powerful and flexible tool for data manipulation and analysis in Python. This step enables easy access and efficient processing of the data.

```
pd.read_csv("insurance_data.csv", encoding='utf-8')
```

B. *Loading the Dataset*

To begin with, the presence of missing values is identified by applying the isnull().sum() function on the dataset. This function returns the count of missing values for each attribute/column.

In this case, the missing values are found in the 'age' and 'region' attribute. To address these missing values, a common approach is to fill them with a suitable value that preserves the integrity of the dataset and minimizes bias in subsequent analyses.

The chosen approach in this project is to fill the missing values in the 'age' attribute with the mean value of the available ages and dropping of records where 'region' is null. This is achieved using the fillna() function with the value parameter set to the mean of the 'age' attribute obtained through the mean() function.

By replacing the missing values with the mean, it ensures that the overall distribution and statistical properties of the 'age' attribute are preserved. This approach is reasonable when the missing value is assumed to be at random, and the mean provides a representative central tendency measure.

After filling the missing values, the dataset is checked again using the isnull().sum() function to confirm that all missing values have been properly handled.

V. Exploratory Data Analysis (EDA)

Exploratory Data Analysis (EDA) is a critical phase in data analysis where the main objective is to gain a deeper understanding of the dataset. It involves employing various statistical and visualization techniques to uncover patterns, trends, and relationships within the data. Through summary statistics, histograms, scatter plots, and correlation analysis, EDA helps identify outliers, assess data distributions, detect potential errors, and explore dependencies between variables. This process aids in making informed decisions regarding data preprocessing, feature engineering, and model selection. EDA serves as a foundation for further analysis, ensuring the data's quality, and extracting meaningful insights for decision-making and problem-solving.

A. *Categorical Variable Analysis*

Categorical variables play a crucial role in insurance claim analysis, as they provide insights into customer characteristics and behavior. Analyzing categorical variables involves examining their distributions, identifying frequent categories, and understanding their relationship with the target variable.

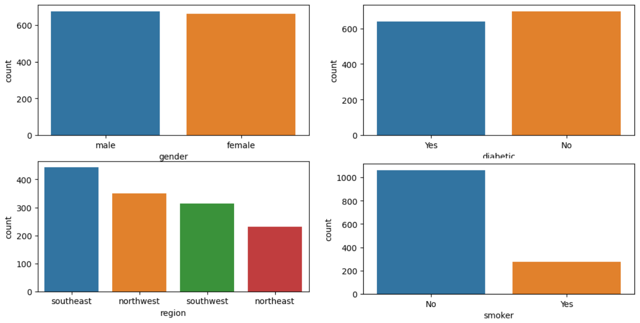

Figure 1

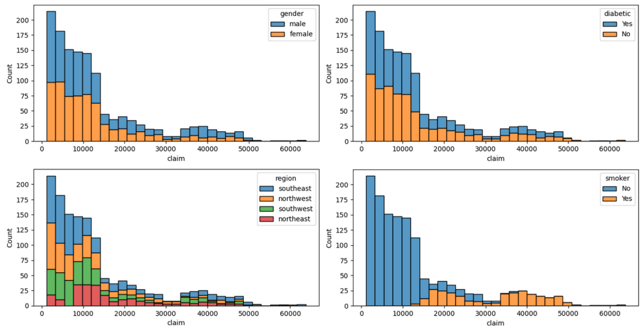

Figure 2

Conclusion about the claims from figure 1 and figure 2:

- 50.5% of beneficiary are male and 49.5 % are female. Approximately same number of male and female beneficiary.
- 20.5% of beneficary are smokers.
- Beneficary are evenly distributed across regions with South East being the most populous one (~27%) with the rest of regions each containing around ~24%
- Most of the beneficiary don't have kid..

B. *Numerical Variable Analysis*

Numerical variables provide valuable insights into claim amounts, policy details, and customer demographics. Analyzing numerical variables involves examining their distributions, identifying outliers, and understanding their relationship with the target variable..

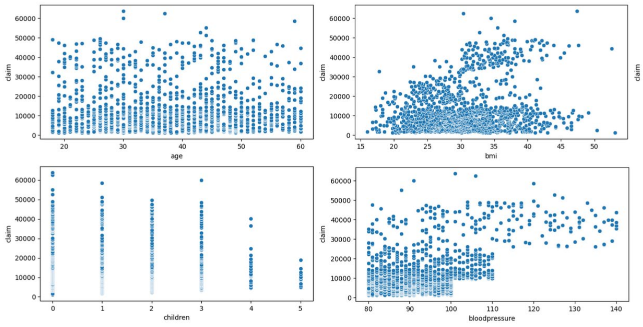

Figure 3

Observations

- Age of primary beneficiary lies approximately between 20 - 65. Average Age is approx. 40. Majority of customer are in range 18- 20's.
- Bmi is normally distributed and Average BMI of beneficiary is 30.This BMI is outside the normal range of BMI. There are lot of outliers at upper end
- Most of the beneficiary have 1, 2 or 3 children.

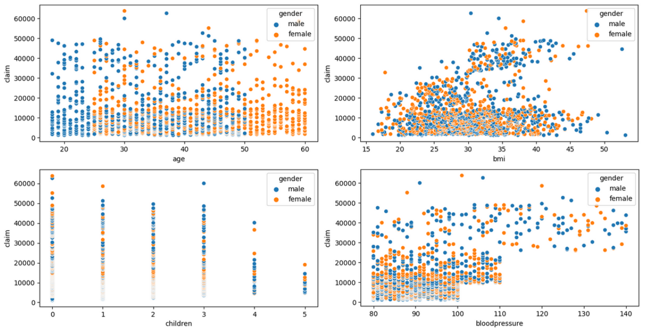

Figure 4

Observations

- Average Age of female beneficiary is slightly higher than male beneficiary
- No of children both male and female beneficary have is same
- BMI of Male policy holder has many outliers and Average BMI of male is slightly higher than female
- Male policy holder has incure more charges to insurance compared to female policy holder. There are lot of outliers in female policy holder

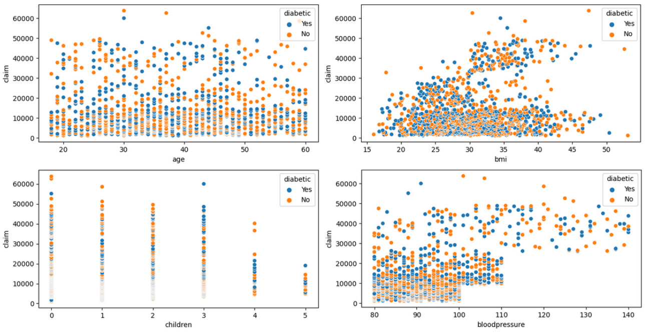

Figure 5

Observations

- People with no diabetics and have 0 or 1 child claim higher insurance.
- People with no diabetics and have blood pressure between 100-120 claim higher insurance.

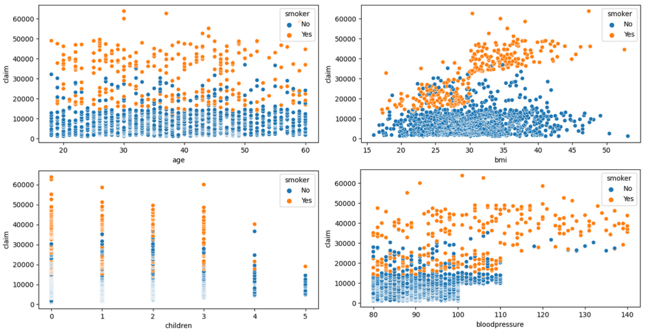

Figure 6

Observations

- Smoker have incurred more cost to insurance than nonsmoker.
- There are outliers in nonsmoker.
- BMI of non-smoker has lot of outliers

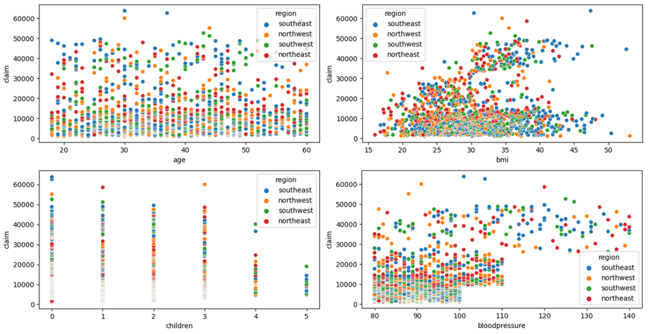

Figure 7

Observations

- Age and number of children across regions is almost same.
- Average Bmi of policy holder from southeast higher compared to other regions
- Charges incurred because of policy holder from southeast is higher compared to other regions

C. *Heat Map*

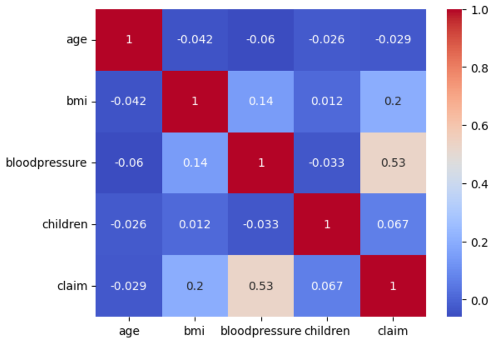

Figure 8

Observations

- Blood pressure has the highest positive correlation score of total claim which is 0.53. Followed by BMI which is 0.2.
- Age and children almost don't have correlation with total claim.

VI. Feature Engineering : One-Hot Encoding

Feature engineering is a crucial step in data preprocessing, and one-hot encoding is a common technique used to transform categorical variables into a numerical representation suitable for machine learning algorithms. With one-hot encoding, each categorical variable is expanded into multiple binary columns, where each column represents a unique category. If a data point belongs to a specific category, the corresponding binary column is marked as 1, and all other columns are marked as 0. This transformation allows the categorical information to be effectively incorporated into machine learning models, as algorithms typically work with numerical inputs. One-hot encoding ensures that each category is treated independently, avoiding any implicit ordering or numerical bias.

In the dataset used, categorical variables such as 'gender,' 'diabetic,' 'smoker,' and 'region' are present. Machine learning algorithms typically require numerical inputs, so these categorical variables need to be transformed into a suitable numerical representation. One-hot encoding is applied to convert each categorical variable into multiple binary columns, where each column represents a unique category. By doing so, the categorical information is effectively encoded as binary values (0 or 1) that can be used by machine learning models.

One-hot encoding ensures that each category is treated independently, without imposing any implicit ordering or numerical bias. It allows the models to consider the categorical variables as separate features, capturing their influence on the insurance claim prediction accurately. It is performed by using the function: pd.get_dummies(df)

VII. Preparation of the Model Test and Training Data

The dataset is divided into two parts: the training set and the testing set. The training set is used to train the machine learning models, allowing them to learn the underlying patterns and relationships present in the data. The testing set, on the other hand, is used to assess the performance of the trained models by evaluating their predictions on unseen data.

The train_test_split() function takes as input the feature matrix (X) and the target variable (y). Additional parameters are provided to specify the proportion of data to be allocated to the training and testing sets. An 8:2 split was specified, indicating that 80% of the data will be used for training and 20% for testing.

The function returns four sets of data: X_train (training features), X_test (testing features), y_train (training target), and y_test (testing target). These sets are further used for model training, evaluation, and comparison.

VIII. Prediction with Various Models

A. *Support Vector Machine*

SVM is a supervised learning algorithm that attempts to find the best possible hyperplane that separates the data points into different classes. In regression, it aims to find the hyperplane that fits the data points with maximum margin while minimizing the error. However, in this case, the obtained metrics indicate poor model performance. The negative R2 score (-0.128) suggests that the model performs worse than a horizontal line.

B. *Random Forest*

Random Forest is an ensemble learning algorithm that combines multiple decision trees to make predictions. It operates by creating a multitude of decision trees and aggregating their predictions to obtain the final prediction. In the project, Random Forest Regression is employed. The obtained metrics show relatively good model performance with a positive R2 score (0.824), indicating that the model explains a significant portion of the variance in the target variable. The RMSE score (5301.17) and MAE score (0.6869) represent the average magnitude of errors made by the model.

C. *Linear Regression*

Linear Regression is a simple yet powerful algorithm that models the linear relationship between the independent variables and the dependent variable. It assumes a linear relationship and estimates the coefficients that best fit the data. The metrics obtained for Linear Regression in the project indicate reasonably good model performance. The positive R2 score (0.718) suggests that the model explains a substantial portion of the variance in the target variable. The RMSE score (6719.71) and MAE score (0.7626) represent the average magnitude of errors made by the model.

The below table shows all the metrics of the algorithms used.

Algorithm	***R2 Score***	***RMSE***	***MAE***
Support Vector Machine	-0.128	13435.258	1.021
Random Forest	0.824	5301.168	0.686
Linear Regression	0.717	6719.712	0.762

Table 1: Algorithm Metrics

IX. CONCLUSION

In conclusion, the Insurance Claim Analysis project successfully developed a robust predictive model for estimating insurance claim amounts. Through thorough data preprocessing and insightful exploratory data analysis, valuable insights were gained regarding the dataset and its features. By implementing Support Vector Machine, Random Forest Regression, and Linear Regression models, accurate predictions of claim amounts were achieved. Notably, the Random Forest Regression model emerged as the top performer, showcasing the highest R2 score and the lowest RMSE and MAE scores. This selected model provides insurance companies with a dependable tool for estimating claim amounts, allowing them to make well-informed decisions, optimize resource allocation, and streamline their operations. Furthermore, areas for future improvement, such as advanced feature engineering techniques and model optimization, were identified to enhance the predictive accuracy of the model. The findings of this project have substantial implications for the insurance industry, offering enhanced capabilities for claim estimation and management.

ACKNOWLEDGMENT

We would like to express our sincere gratitude to our supervisor Dr. Jayita Saha for their invaluable guidance and support throughout our study. We are thankful to Dayananda Sagar University for providing us with the necessary resources and facilities. We also extend our appreciation to the authors of the open-source libraries and frameworks used in this project. Lastly, we would like to thank the participants whose data contributed to this analysis. Their contributions were essential to the project's success.

REFERENCES

[1] Sumit Kumar Shukla. (2021). insurance, Version 1. Retrieved May 10, 2023 from https://www.kaggle.com/datasets/thedevastator/insurance-claim-analysis-demographic-and-health.

[2] M. A. Hearst, S. T. Dumais, E. Osuna, J. Platt and B. Scholkopf, "Support vector machines," in IEEE Intelligent Systems and their Applications, vol. 13, no. 4, pp. 18-28, July-Aug. 1998, doi: 10.1109/5254.708428.

[3] J. K. Jaiswal and R. Samikannu, "Application of Random Forest Algorithm on Feature Subset Selection and Classification and Regression," 2017 World Congress on Computing and Communication Technologies (WCCCT), Tiruchirappalli, India, 2017, pp. 65-68, doi: 10.1109/WCCCT.2016.25.

[4] M. Huang, "Theory and Implementation of linear regression," 2020 International Conference on Computer Vision, Image and Deep Learning (CVIDL), Chongqing, China, 2020, pp. 210-217, doi: 10.1109/CVIDL51233.2020.00-99.

[5] A. S. Rao, B. V. Vardhan and H. Shaik, "Role of Exploratory Data Analysis in Data Science," 2021 6th International Conference on Communication and Electronics Systems (ICCES), Coimbatre, India, 2021, pp. 1457-1461, doi: 10.1109/ICCES51350.2021.9488986.