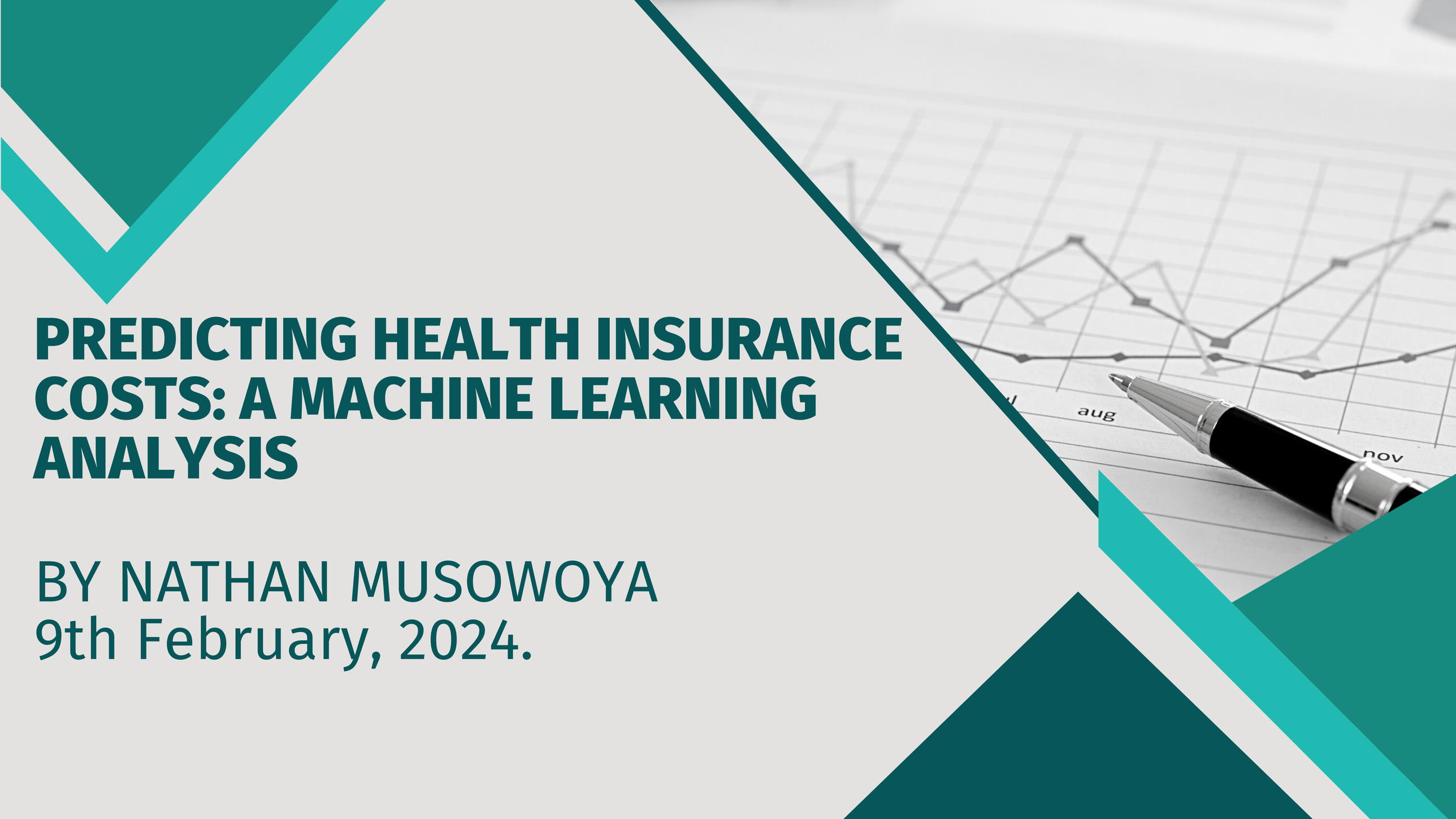
PREDICTING HEALTH INSURANCE COSTS: A MACHINE LEARNING ANALYSIS



BY NATHAN MUSOWOYA
9th February, 2024.

Table Of Content

INTRODUCTION

Forecasting consumer healthcare spending enhances accountability within the industry. Despite extensive patient, illness, and diagnosis data, insights into patient costs are often underutilized (Somal (2020)). Medical system goals include quality care for many, though costs remain vital. Practitioners and support staff require compensation, making expensive treatments unaffordable for individuals. Insurance plans spread expenses to maintain necessary resources. Machine learning aids healthcare through rapid claim processing, personalized plans, fraud detection, and improved drug discovery, potentially lowering costs (Bhattacharya et al. (2020)). Data mining supports the creation of cost-effective, targeted medications. This research predicts insurance costs to help companies align spending with customer demographics, health history, and lifestyle choices. Such early cost estimates could support informed patient decisions. This builds on work by Iqbal et al. (2021) exploring Generalized Linear Regression for prediction and uses RMSE and R2 scores for model evaluation.

Research Question and objective

Research Question:

How accurately we can predict Health Insurance premiums based on customers medical background and yearly premium expense with the help of Linear Regression and Regularization techniques to assist healthcare personnel in spending more time delivering appropriate treatment, lowering burnout among medical experts?

OBJECTIVES

- Gather the dataset and perform Data Pre-processing for model building
- Implement Linear Regression Model with variations
- Evaluate the results and find out which model is best for predicting Insurance claim

The main goal of this idea is to assist the healthcare business in forecasting the cost of an insurance claim for a specific person and expediting the health insurance procedure. Section 2 of this discusses the Related work on the topic of Insurance Premiums and Linear Regresssion Model. Section 3 provides the brief explanation of Methodology. Section 4 defines the workflow of the Data analysis. Section 5 provides information of Implementation of the Machine Learning Model. Evaluation of all the models are carried out in Section 6. Conclusion and Future work is discussion in Section 7.

RELATED WORK

Christobel and Subramanian (2022) proposed a model for predicting health insurance costs, finding polynomial regression effective with 88% accuracy. To detect fraudulent medical insurance claims, Bauder et al. (2016) analyzed billing data through a multinomial Naive Bayes algorithm showing strong F-scores. For similar detection and premium calculations, Kowshalya and Nandhini (2018) compared Random Forest, Naive Bayes, and J48 classifiers on financial and personal data, with Random Forest reaching 99.41% accuracy under certain conditions.

Mustika et al. (2019) employed XGBoost to evaluate life insurance risk, reporting 60.7% accuracy. Morid et al. (2017) reviewed prediction literature, finding supervised learning techniques generally more accurate, with Gradient Boosting often offering the best performance. Other studies utilized regression (Panay et al. 2019; Freyder 2016), ANNs (Kaushik et al. 2022; El Bouanani et al. 2022), and combined techniques (Agarwal and Tripathi 2022). Overall, diverse data and algorithmic approaches continue to drive research in accurate health insurance cost prediction.

METHODOLOGY

The Cross Industry Standard Process for Data Mining (CRISP-DM) includes a hierarch ical and iterative process model, as well as a framework that can be expanded using a generic-to-specific approach. It begins with six phases and then delves further into general and then specialized jobs stated in Figure 1. The approach is adaptable to accommodate a range of formality levels that may vary in DM projects of different sizes and levels of complexity.(Niaksu (2015)

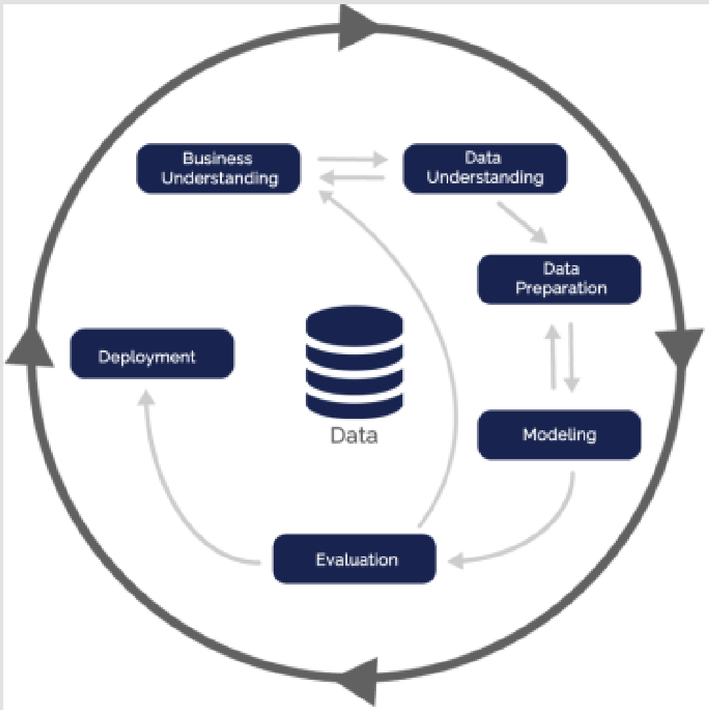


Figure 1: Methodology

BUSINESS UNDERSTANDING

Predicting health insurance premiums using machine learning algorithms is still a topic that needs further research and exploration in the medical care industry. Even as the healthcare sector digitizes more and more, enormous amounts of data will inevitably be created and gathered. The health insurance industry may achieve a number of objectives with the help of AI and machine learning. Personalized health insurance plans, more in expensive insurance options, the capacity to spot insurance fraud, advances in medication discovery, and quicker claim processing are just a few of the critical elements.

DATA UNDERSTANDING

This stage begins with data collection and getting familiar with the data. The dataset was extracted Kaggle data repository. The data consists of 15000 rows and 14 variables.

This data will be divided into 70:30 ratio for training and testing the model. The data set consists of customer's personal medical records and details taken by the Insurance company for initializing their premium amount.

1www.otaris.com

2https://www.kaggle.com/datasets/sureshgupta/health-insurance-data-set

Variable	Description	Datatype
age	Age of the policyholder	Numeric
sex	Gender of policyholder	Categoric
weight	Weight of the policyholder	Numeric
bmi	Body mass index	Numeric
no_of_dependents	Number of dependent persons on the policyholder	Numeric
smoker	Indicates policyholder is a smoker or a non-smoker	Categoric
bloodpressure	Bloodpressure reading of policyholder	Numeric
diabetes	Indicates policyholder suffers from diabetes or not	Categoric
regular_ex	A policyholder regularly excercises or not	Categoric
job_title	Job profile of the policyholder	Categoric
city	The city in which the policyholder resides	Categoric
state	The state in which the policyholder resides	Categoric
hereditary_diseases	A policyholder suffering from hereditary diseases or not	Categoric
claim	The amount claimed by the policyholder	Numeric

Figure 2: Data Description

	age	weight	bmi	no_of_dependents	bloodpressure	claim
count	14604.000000	15000.000000	14044.000000	15000.000000	15000.000000	15000.000000
mean	39.547521	64.909600	30.266413	1.129733	68.650133	13401.437620
std	14.015966	13.701935	6.122950	1.228469	19.418515	12148.239619
min	18.000000	34.000000	16.000000	0.000000	0.000000	1121.900000
25%	27.000000	54.000000	25.700000	0.000000	64.000000	4846.900000
50%	40.000000	63.000000	29.400000	1.000000	71.000000	9545.650000
75%	52.000000	76.000000	34.400000	2.000000	80.000000	16519.125000
max	64.000000	95.000000	53.100000	5.000000	122.000000	63770.400000

	sex	hereditary diseases	smoker	city	state	diabetes	regular_ex	job_title
count	15000	15000	15000	15000	15000	15000	15000	15000
unique	2	10	2	91	35	2	2	35
top	female	No Disease	0	New Orleans	California	1	0	Student
freq	7652	13998	12028	302	2003	11655	11638	1320

Figure 3: Descriptive Statistics of all features

The output in Figure 3 illustrates the summary statistics of all the numeric variables like the mean, median(50%), minimum, and maximum values, along with the standard deviation. Note, the average age of a policyholder claiming the insurance is 39 years. The claim amount is between 1121 to 63770. Here the mean BMI of a policyholder is 30 (the healthy BMI range is between 16 to 24.9) and the average weight is 64.

	Total	Percentage of Missing Values
bmi	956	6.373333
age	396	2.640000
sex	0	0.000000
weight	0	0.000000
hereditary diseases	0	0.000000
no_of_dependents	0	0.000000
smoker	0	0.000000
city	0	0.000000
state	0	0.000000
bloodpressure	0	0.000000
diabetes	0	0.000000
regular_ex	0	0.000000
job_title	0	0.000000
claim	0	0.000000

	Total	Percentage of Missing Values
age	0	0.000000
sex	0	0.000000
weight	0	0.000000
bmi	0	0.000000
hereditary diseases	0	0.000000
no_of_dependents	0	0.000000
smoker	0	0.000000
city	0	0.000000
state	0	0.000000
bloodpressure	0	0.000000
diabetes	0	0.000000
regular_ex	0	0.000000
job_title	0	0.000000
claim	0	0.000000

Figure 4: Dealing with Missing Values

If we observe the count of all the variables in Figure 3, there is less count for variable age and BMI than other variables. So we can say that there are missing values in these variables. The missing values are handled by replacing them with the mean of the variable 'age' and 'bmi' as can be seen in Figure 4. Also, the minimum blood pressure is zero, which is invalid. Hence, the zeros were replaced by the median of the values in blood pressure.

FEATURE ENGINEERING

There is a total of 91 cities in the 'city' variable. A new variable 'Region' is created with the 'city' variable. These 91 cities are divided into 4 regions named 'North-East', 'Southern', 'Mid-West', and 'West'.In Figure 6, it can be observed that for both males and females insurance premium claims are increasing with the increase in age. The distri bution of claims between the two categories, 'smoker'(1) and 'non-smoker'(0), are distinct enough to take smokers as a potentially good predictor of the claim amount. The distri bution of claims between the two categories, 'smoker'(1) and 'non-smoker'(0), are distinct enough to take smokers as a potentially good predictor of the claim amount. We can see that a 'non-smoker' has a median claim amount of around 10000 while a 'smoker' has a median claim of 40000.

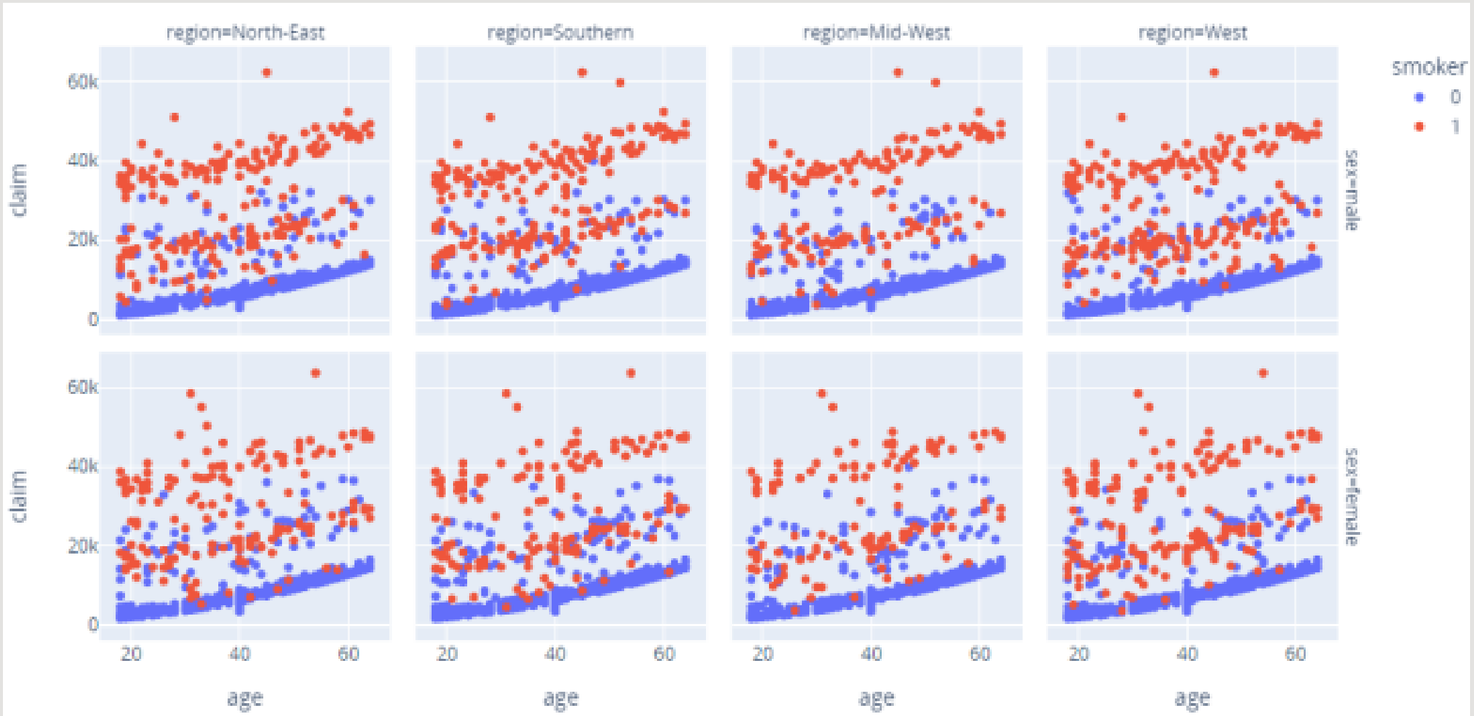


Figure 6: Analysis of new Feature "Region" vs claim, age and smoker

In Figure 7, it can be observed that the insurance claim is significantly increasing with an increase in BMI of the customers. The healthy range of BMI is considered to be between 18.5 to 24.9 (Klein et al. (2007)). When a person's BMI is above 30 and they smoke, their insurance rates are often over $40,000, according to observations. Customers who smoke but have a healthy BMI pay less in premiums.

Figure 7: Analysis of new Feature "Region" vs claim, bmi and smoker

MODELLING

Linear Regression: A statistical method called linear regression is used to simulate the linear connection between a dependent variable and one or more independent variables.
It is based on the linear equation, which states the following relationship between the dependent variable (Y) and the independent variables (X):
$Y = b0 + b1 * X1 + b2 * X2 + ... + bn * Xn$ (1)
Where b0 is the intercept, and b1, b2, ..., bn are the regression coefficients. By reducing the total of squared deviations between the observed and predicted values of the dependent variable, linear regression calculates the values of the regression coefficients. The ensuing regression equation is known as the least squares regression equation, and the approach used is known as the least squares technique.

REGULARIZATION:

In order to minimize the adjusted loss function and avoid over-fitting or under-fitting, regularization refers to methods for calibrating machine learning models. Regularization
allows us to properly adapt our machine learning model to a specific test set, hence lowering the mistakes in the test set.

LASSO AND RIDGE REGRESSION:

Lasso and Ridge regression are both types of regularized linear regression, which are used to address the problem of multicollinearity in linear regression. Multicollinearity occurs when the independent variables in a regression model are highly correlated and can lead to unstable and unreliable estimates of the regression coefficients. Lasso and Ridge regression both address multicollinearity by adding a penalty term to the cost function that is used to train the model. This penalty term, called the regularization term, is designed to penalize models with large coefficients, which can help to reduce the impact of multicollinearity and improve the stability of the model.
The Lasso and Ridge regression equations are similar to the linear regression equation but include the regularization term in addition to the terms for the independent variables.
The Lasso regression equation is of the form:
$Y = b0 + b1 * X1 + b2 * X2 + ... + bn * Xn + lambda *$
$X | bi |$ (2)
Where lambda is the regularization parameter, which controls the strength of the regularization, and the summation is over all the coefficients of the independent variables.

ELASTIC NET REGRESSION:

Elastic Net regression is a type of regularized regression method that combines the penalties of both L1 and L2 regularization. It is called ”Elastic Net” because it can be seen as a combination of both L1 and L2 regularization, which are typically represented by a ”net”-like structure. The regularization term for Elastic Net regression is defined as follows:
$\lambda * (\alpha * L1\,norm + (1 - \alpha) * L2\,norm)$ (3)
where λ is the regularization parameter, is a mixing parameter that determines the balance between L1 and L2 regularization, and L1 norm and L2 norm are the L1 and L2 norms of the model weights, respectively.
In other words, Elastic Net regression is a linear regression model with a regularization term that penalizes both the L1 and L2 norms of the model weights. This can help to prevent overfitting and improve the generalization performance of the model.

EVALUATION

In this final stage, the performance of a model in forecasting Medical Insurance Claim cost is compared with other models based on the evaluation metrics RMSE, R-Squared, and Adjusted R-Squared.

• RMSE(Root Mean Square Error): Root Mean Square Error (RMSE) is a standard way to measure the error of a model in predicting quantitative data. It indicates the absolute fit of the model to the data.

• R-Squared: A statistical fit indicator known as R-Squared quantifies how much variance in a dependent variable is explained by one or more independent variables in a regression model.

• Adjusted R-Squared: By taking into account the influence of extra independent factors that have the propensity to distort the outcomes of R-squared measure ments, adjusted R-squared, a modified form of R-squared, increases accuracy and dependability.

The metrics used for the analysis of the machine learning model give the complete picture of their performance over the train and test dataset. The difference between R2 and Adjusted R2 is that both explain how much variance in the dependent variable is explained by the independent variable but Adjusted R2 determines whether any addition of the independent variable decrease the accuracy of the model or not. If there is less difference in R2 and Adjusted R2 then it states that all the independent variables/predictors are significant. Root mean square error is useful in the scenario of forecasting the results based on historical data. It gives the standard deviation of residuals, the measure which explains how far the predicted values are from the actual values. P-value gives the signi ficance of the regression model. Durbin Watson explains if there is any auto-correlation between the predictor/independent variables.

DEPLOYMENT

The project's ultimate goal is not to build modeling; rather, this final stage is putting its research and analysis into a written format that can be easily read. Despite the fact that modeling is meant to provide additional detail to the data, this knowledge still has to be arranged and presented in such a manner that customers can utilize it. The likelihood cost of claims for the insurance industry can only be effectively reduced by disclosing the forecast to the decision maker.

DESIGN SPECIFICATION

Following data pre-processing, machine learning techniques will be implemented. The implementation of multiple linear regression will take many factors into account. with or without the interaction of significant factors, deleting insignificant variables, and de pendent variables that have been log-transformed. The performance will be improved by hyperparameter optimization. The following procedure will be used by these algorithms

in Figure 8.

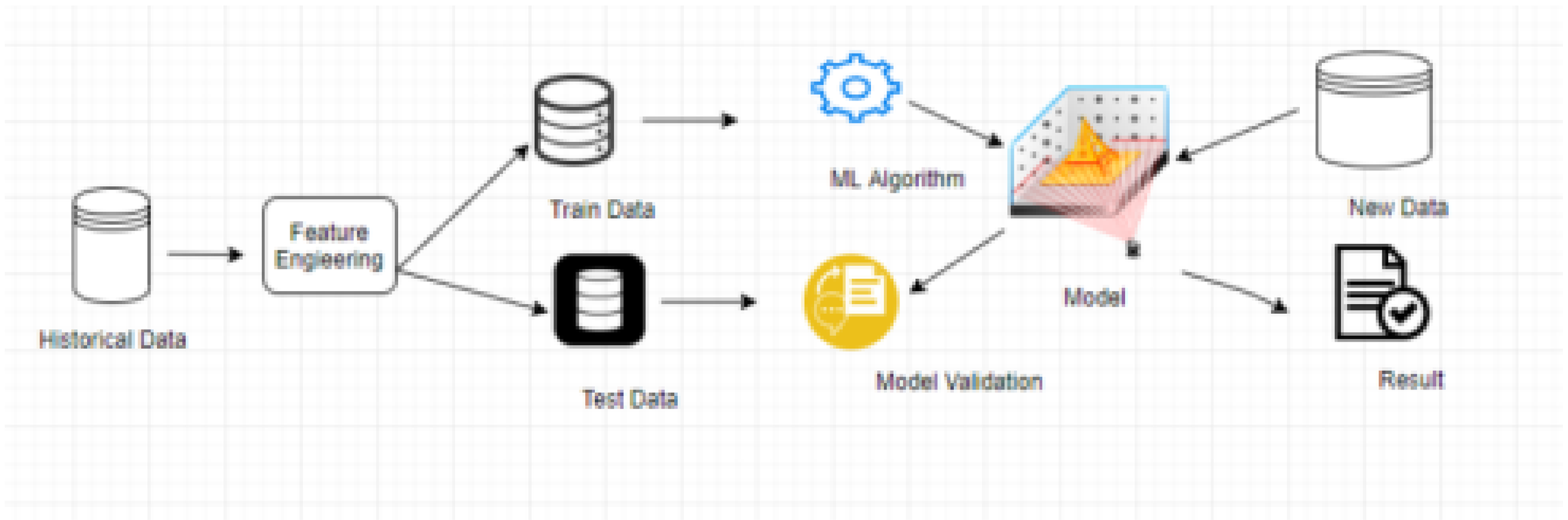


Figure 8: Workflow diagram

IMPLEMENTATION

All the data analysis is done in Python language on Jupyter notebook. The Data preparation is done with the application of the mean, and median of the variables. The models are implemented with the Ordinary Least Square method. The ordinary least squares (OLS) method is a linear regression analysis method that estimates the parameters of a linear regression model by minimizing the sum of the squares of the differences between the observed responses in the dataset and the responses predicted by the linear model. The goal of OLS is to find the line that best fits the data.(Whatley (2022)) The output gives are R2 score of the model which depicts the value, what percentage variance of the target variable is explained by the model. The linear Regression model is explored in different scenarios. In order to determine if a hyper-parameter adjustment will improve the model's accuracy, regularization techniques were investigated.

EVALUATION

This section uses machine learning techniques to evaluate the results of the model built. All the processes of evaluation are conducted in Python to run the Machine learning algorithms. The metrics used to evaluate the regression models are R2 score, Adjusted R2 score, Root Mean Square Error, P-value, and Durbin Watson test.

When using a multiple linear regression model, the dependent variable must be nu merical and have a minimum of two independent variables. Let's say that the function of the dependent variable translates to other variables and some random noise.

The model creates data using the format $y = \beta 0 + \beta 1x1 + \beta 2x2 + \ldots + \beta kxk + \epsilon$ where the dependent variable is y, the independent variables are $\beta 0 + \beta 1x1 + \beta 2x2 + \ldots + \beta kxk$, is random noise (error) and βi is the contribution value of the independent variables. When all of the independent variables are zero, the dependent variable's value is represented by the y-intercept.

EXPERIMENT 1: MULTIPLE LINEAR REGRESSION – FULL MODEL – WITH SQUARE ROOT TRANSFORMED DEPENDENT VARIABLE

In this section, a full model with linear regression has been built using OLS (Ordinary Least Square) technique. The full model indicates that all the independent variables have been considered that are present in the dataset. Independent Variable: age, weight, bmi, no of dependents, blood pressure, sex, hered itary diseases, smoker, state, diabetes, regular ex, job title, region Dependent Variable: claim

OLS Regression Results

Dep. Variable:	sqrt_claim	R-squared:	0.763
Model:	OLS	Adj. R-squared:	0.761
Method:	Least Squares	F-statistic:	388.0
Date:	Wed, 14 Dec 2022	Prob (F-statistic):	0.00
Time:	10:38:50	Log-Likelihood:	-46637.
No. Observations:	10238	AIC:	9.345e+04
Df Residuals:	10148	BIC:	9.410e+04
Df Model:	89		
Covariance Type:	nonrobust		
Omnibus:	3339.582	Durbin-Watson:	2.038
Prob(Omnibus):	0.000	Jarque-Bera (JB):	11194.399
Skew:	1.655	Prob(JB):	0.00
Kurtosis:	6.909	Cond. No.	7.06e+03

Model_Name	R-Squared	Adj. R-Squared	RMSE
Linreg full model with sqrt of target variable	0.763457	0.761383	5881.709984

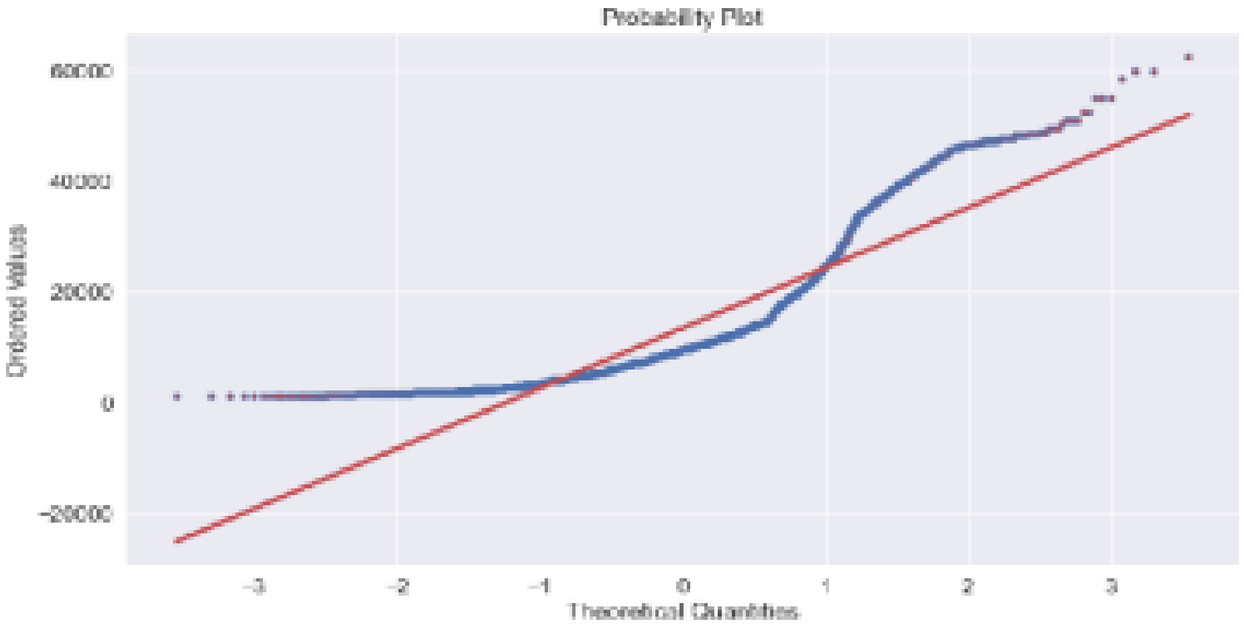


Figure 9: Results of Experiment 1

From the results in Figure 9, This model explains 76.3% of the variation in the de pendent variable 'claim'. The Durbin-Watson test statistics is 2.029 and indicates that there is no autocorrelation. Condition Number 7.06e+03 suggests that there is severe collinearity. The Q-Q plot depicts the predicted claim point in blue color and shows how close it is to the actual points ie. the best-fit model line.

EXPERIMENT 2: MULTIPLE LINEAR REGRESSION - FULL MODEL - WITH SQUARE ROOT TRANSFORMED DEPENDENT VARIABLE

In this section, any kind of transformation on the dependent variable is not considered, the dependent variable 'claim' is used as it is.

OLS Regression Results

Dep. Variable:	claim	R-squared:	0.758
Model:	OLS	Adj. R-squared:	0.755
Method:	Least Squares	F-statistic:	350.3
Date:	Wed, 14 Dec 2022	Prob (F-statistic):	0.00
Time:	11:03:59	Log-Likelihood:	-1.0345e+05
No. Observations:	10238	AIC:	2.071e+05
Df Residuals:	10148	BIC:	2.077e+05
Df Model:	89		
Covariance Type:	nonrobust		

Omnibus:	2964.784	Durbin-Watson:	2.033
Prob(Omnibus):	0.000	Jarque-Bera (JB):	10475.005
Skew:	1.435	Prob(JB):	0.00
Kurtosis:	7.040	Cond. No.	7.06e+03

Model_Name	R-Squared	Adj. R-Squared	RMSE
Linreg full model without sqrt of target variable	0.757544	0.755417	5858.244928

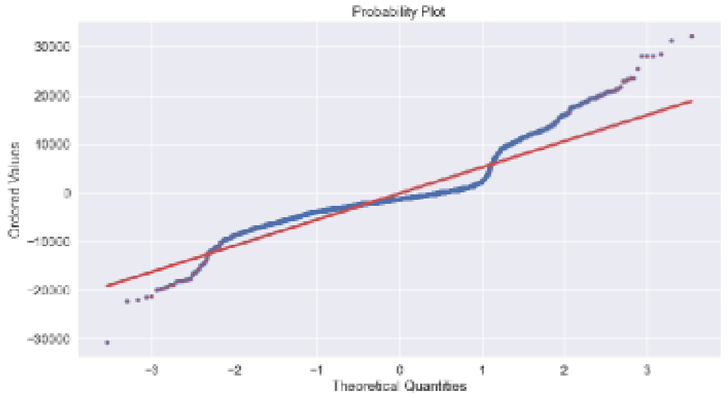
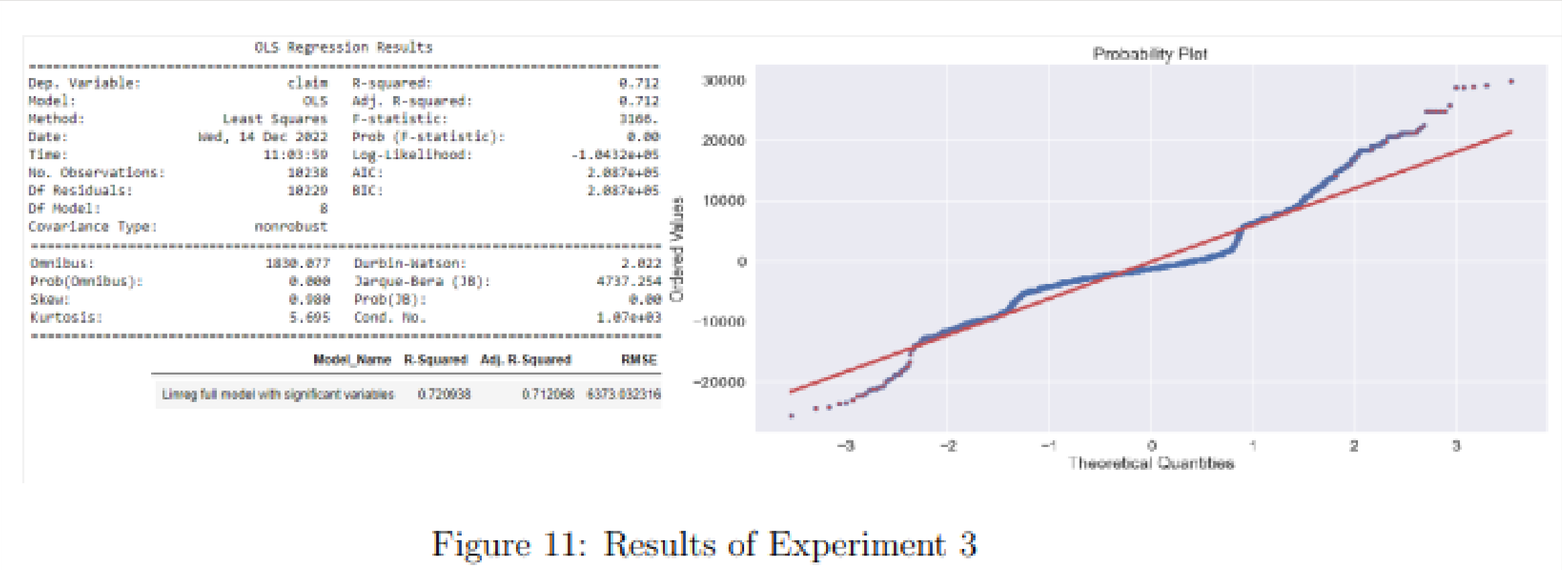


Figure 10: Results of Experiment 2

This model explains 75.8% of the variation in the dependent variable 'claim'. The Durbin-Watson test statistics is 2.033 and indicates that there is no autocorrelation. Condition Number 7.06e+03 suggests that there is severe collinearity. The Q-Q plot depicts the predicted claim point in blue color and shows how close it is to the actual points ie. the best-fit model line. On comparing the above models in Experiment 1 and 2, it is seen that the R-squared and the Adjusted R-squared value for the model considering square root transformation of the variable 'claim' is lower than the other model. And, the RMSE value of the model without considering the square root transformation is considerably lower. So, we continue with variable 'claim' as it is, instead of opting for log transformation

EXPERIMENT 3: LINEAR REGRESSION WITH SIGNIFICANT VARIABLE

After comparing the P-value in Experiment 2, it was found that 'sex', 'job title', 'region', 'state', and 'hereditary diseases' features are insignificant to predict the dependent variable 'claim' as they have P-value greater than 0.05. Occam's razor is a principle to explain the phenomena by the simplest hypothesis possible. The last model where the insignificant variables are removed is performing very close to the other models in spite of having a lesser number of variables. Using Occam's razor principle, the model is accepted in which we consider the model with significant variables.(Blumer et al. (1987)) This model in explains 71.2% of the variation in the dependent variable 'claim'. The Durbin-Watson test statistics is 2.022 and indicates that there is no autocorrelation. Condition Number 1.07e+03 suggests that there is severe collinearity. Since the accuracy is low for this Experiment than Experiment 2 and. This model is rejected.



OLS Regression Results

Dep. Variable:	claim	R-squared:	0.712
Model:	OLS	Adj. R-squared:	0.712
Method:	Least Squares	F-statistic:	3168.
Date:	Wed, 14 Dec 2022	Prob (F-statistic):	0.00
Time:	11:03:59	Log-Likelihood:	-1.0432e+05
No. Observations:	10238	AIC:	2.087e+05
Df Residuals:	10229	BIC:	2.087e+05
Df Model:	8		
Covariance Type:	nonrobust		
Omnibus:	1830.877	Durbin-Watson:	2.022
Prob(Omnibus):	0.000	Jarque-Bera (JB):	4737.254
Skew:	0.980	Prob(JB):	0.00
Kurtosis:	5.695	Cond. No.	1.07e+03

Model_Name	R-Squared	Adj. R-Squared	RMSE
Linreg full model with significant variables	0.720938	0.712068	6373.032316

Figure 11: Results of Experiment 3

EXPERIMENT 4: ELASTIC NET REGRESSION

In this Experiment, the elastic net is a regularized regression method that linearly com bines the L1 and L2 penalties of the lasso and ridge methods. The model is trained using both L1 L2(Lasso and Ridge Regression) which allows learning of sparse model where few entries are zero similar to Lasso and also maintaining the regularization properties similar to ridge regression. The parameter for Elastic Net regression is selected with the help of GridsearchCV. It is supposed to be between 0 to 1. For Ridge Regression the L1 ratio is taken as 1 and for Lasso Regression it is 0. The best parameter for L1 ratio for Elastic Net Regression is 0.2 according to the GridSearchCV technique with 10 cross validations. The alpha value is set to 0.0001 because it gives the minimum RMSE value after applying the model. The R2 value of this experiment is 75.5% which is similar to

OLS Regression Results

Dep. Variable:	claim	R-squared:	0.758
Model:	OLS	Adj. R-squared:	0.755
Method:	Least Squares	F-statistic:	356.2
Date:	Wed, 14 Dec 2022	Prob (F-statistic):	0.00
Time:	11:06:24	Log-Likelihood:	-7252.3
No. Observations:	10238	AIC:	1.470e+04
Df Residuals:	10148	BIC:	1.536e+04
Df Model:	89		
Covariance Type:	nonrobust		
Omnibus:	2973.220	Durbin-Watson:	2.033
Prob(Omnibus):	0.000	Jarque-Bera (JB):	18518.032
Skew:	1.439	Prob(JB):	0.00
Kurtosis:	7.847	Cond. No.	115.

Model_Name	R-Squared	Adj. R-Squared	RMSE
Elastic net Regression	0.757537	0.755411	5856.444785

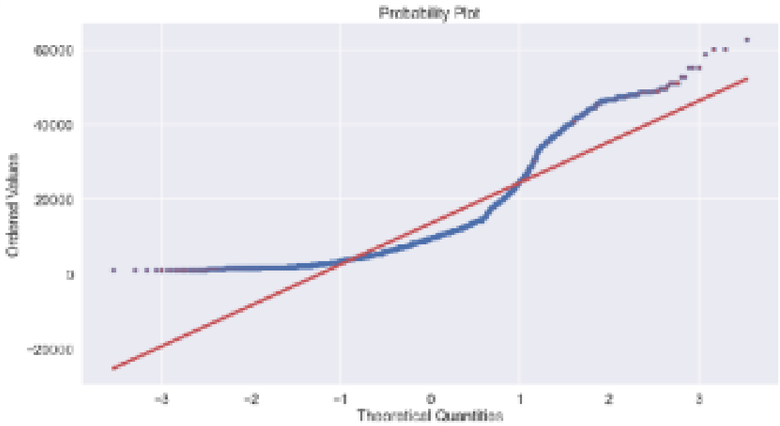
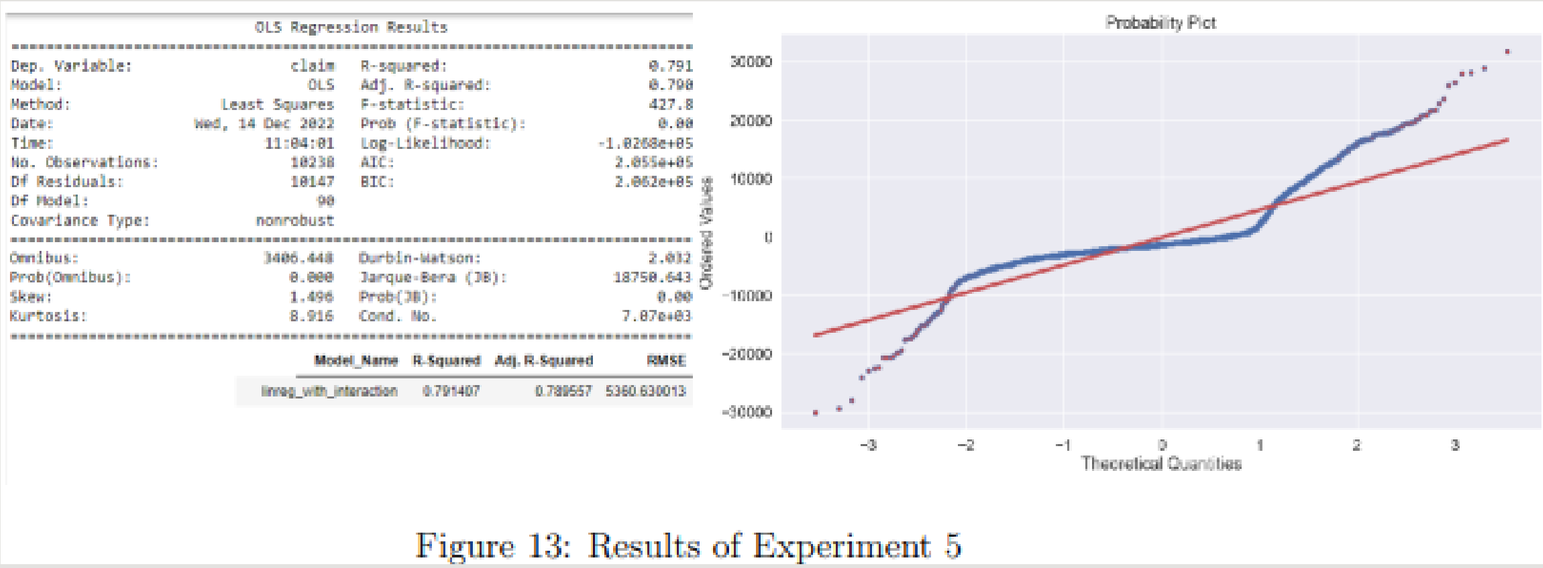


Figure 15: Results of Experiment 7

Ridge and Lasso Regression. Overall, the regularization technique doesn’t improve the model’s accuracy.

EXPERIMENT 5:LINEAR REGRESSION WITH INTERACTION

In this Experiment in Figure 13, a new variable is introduced with the interaction of BMI and smokers. The interaction of 2 variables provides a new possibility of predicting the dependent variable 'claim'. Any 2 significant variables can be used in this case.
Linear regression with interaction terms can be useful in situations where the relationship between the dependent and independent variables is not well-represented by a simple linear model. By adding interaction terms to the model, you can capture non-linear relationships and potentially improve the model's fit and predictive ability. It helps to improve the fit and predictive ability of your model, especially in situations where the relationship between the dependent and independent variables is more complex than a simple linear relationship.



OLS Regression Results

Dep. Variable:	claim	R-squared:	0.791
Model:	OLS	Adj. R-squared:	0.790
Method:	Least Squares	F-statistic:	427.8
Date:	Wed, 14 Dec 2022	Prob (F-statistic):	0.00
Time:	11:04:01	Log-Likelihood:	-1.0268e+05
No. Observations:	10238	AIC:	2.055e+05
Df Residuals:	10147	BIC:	2.062e+05
Df Model:	90		
Covariance Type:	nonrobust		

Omnibus:	3486.448	Durbin-Watson:	2.032
Prob(Omnibus):	0.000	Jarque-Bera (JB):	18750.643
Skew:	1.496	Prob(JB):	0.00
Kurtosis:	8.916	Cond. No.	7.07e+03

	Model_Name	R-Squared	Adj. R-Squared	RMSE
	linreg_with_interaction	0.791407	0.789557	5360.630013

Figure 13: Results of Experiment 5

This model explains 79.1% of the variation in the dependent variable 'claim'. This is the best model so far with RMSE of 5360. This value of RMSE states that it has less residual values. In the Q-Q plot in Figure 13, from quantile -2 to 2, the data points pass from the best line fit. That means all the values between this distribution are more or less correctly predicted. In the range outside -3 and 3 quartile, the data points are drifting apart but that is less in number and can be ignored in terms of a large dataset going across 20000 rows and more. The Durbin-Watson test statistics is 2.032 and indicates that the is no autocorrelation. Condition Number 7070 suggests that there is severe collinearity. The collinearity is likely to increase because of the interaction effect.

DISCUSSION

Table1 compares the R2, Adjusted R2 Root Mean Square Error(RMSE) of all the 7 models in the variation of multiple Linear Regression models. Linear Regression with the Interaction of 2 variables turns out to be the best model for the given data of customers

Table 1: Results of all the Models Applied

Model Name	R-Squared	Adj.R-Squared	RMSE
Linear Reg with Sq. root of target variable	0.763457	0.761383	5881.709984
Linear Reg without Sq. root of target variable	0.757544	0.755417	5858.244928
Linear Reg with Significant variables	0.720938	0.712068	6373.032316
Linear Reg with Scaled Significant variables	0.749112	0.748621	5942.347211
Linear Reg with Interaction of 2 variables	0.791407	0.789557	5360.630013
Ridge Regression	0.757478	0.755415	5858.099240
Lasso Regression	0.757478	0.755352	5852.182282
Ridge Regression	0.757537	0.755411	5856.444785

medical records. The model accuracy is 79.1% on the test data set with an RMSE of 5360.63. For other models, the accuracy is more or less similar to each other. Among these models, the transformation of the target variable has performed well with an accuracy of 76.3%. Overall, the normalization of the dataset has not improved the model. Regularization of the model generally improves the model by penalizing the insignificant variables to reduce the complexity of the model. But in this case, it is similar to the normal Linear Regression model with just 75% accuracy. Hence, The model Linear Regression with Interaction of 2 variables can be used by an Insurance company to predict the claim for future customers.

Conclusion and Future Work

This study aims to advance healthcare by investigating different Linear Regression methods to help healthcare professionals and reduce burnout within the system. Insurance companies play a vital role in ensuring patients receive quality care while covering necessary costs. Machine learning can streamline insurance claim processing. The best model found in this research used Linear Regression with Interaction of 2 variables, achieving 79.1% accuracy. This type of forecasting benefits insurers, ensuring appropriate healthcare premiums, and builds trust between policyholders and insurers. Future analysis of insurance company characteristics using internal and external data could offer predictive value. Understanding consumer behavior helps insurers tailor products and pricing. Additional work could include the Stochastic Gradient Descent model for hyperparameter tuning. While the dataset contained biases, such as greater representation of those without hereditary diseases, this research offers foundational steps to assist doctors and insurers with treatment options. Analyzing larger, unbiased datasets with regularization techniques could further improve insurance premium prediction and risk management models.

Question 1

Which insurance policy would be appropriate and desired by various insurer types?

Question 2

How much of insurance cost should be based on a particular patient and behavior?

Question 3

How smoking cigarettes can affect the cost of insurance?

Question 4

How can insurance providers and insurers establish strong, trusting relationships?

ACKNOWLEDGEMENTS

My sincere gratitude goes to my parents, Tilson Musowoya and Mirriam Nzima Musowoya, for their unwavering support, and to my colleagues for their invaluable assistance in brainstorming techniques and identifying errors. This report would not have been possible without their collective contributions.

REFERENCES

Agarwal, D. and Tripathi, K. (2022). A framework for structural damage detection system in automobiles for flexible insurance claim using iot and machine learning, 2022 International Mobile and Embedded Technology Conference (MECON), IEEE, pp. 5–8.

Bhardwaj, N., Anand, R. and Gupta, A. D. (2020). Health insurance amount prediction international journal of engineering research & technology, (IJERT) 9(05).

Christobel, Y. A. and Subramanian, S. (2022). An empirical study of machine learning regression models to predict health insurance cost, Webology (ISSN: 1735-188X) 19(2)

Freyder, C. (2016). Using linear regression and mixed models to predict health care costs after an inpatient event, PhD thesis, University of Pittsburgh

Agarwal, D. and Tripathi, K. (2022). A framework for structural damage detection system in automobiles for flexible insurance claim using iot and machine learning, 2022 International Mobile and Embedded Technology Conference (MECON), IEEE, pp. 5–8.

Goundar, S., Prakash, S., Sadal, P. and Bhardwaj, A. (2020). Health insurance claim prediction using artificial neural networks, International Journal of System Dynamics Applications (IJSDA) 9(3): 40–57.

Kaltsounidis, A. and Karali, I. (2020). Dempster-shafer theory: ηow constraint program ming can help, International Conference on Information Processing and Management of Uncertainty in Knowledge-Based Systems, Springer, pp. 354–367.

Kaushik, K., Bhardwaj, A., Dwivedi, A. D. and Singh, R. (2022). Machine learning-based regression framework to predict health insurance premiums, International Journal of Environmental Research and Public Health 19(13): 7898.

Kowshalya, G. and Nandhini, M. (2018). Predicting fraudulent claims in automobile insurance, 2018 Second International Conference on Inventive Communication and Computational Technologies (ICICCT), pp. 1338–1343.

Morid, M. A., Kawamoto, K., Ault, T., Dorius, J. and Abdelrahman, S. (2017). Su pervised learning methods for predicting healthcare costs: systematic literature review and empirical evaluation, AMIA Annual Symposium Proceedings, Vol. 2017, American Medical Informatics Association, p. 1312.

Petit-Renaud, S. and Denœux, T. (2004). Nonparametric regression analysis of uncer tain and imprecise data using belief functions, International Journal of Approximate Reasoning 35(1): 1–28.

Sommers, B. D. (2020). Health insurance coverage: what comes after the aca? an examination of the major gaps in health insurance coverage and access to care that remain ten years after the affordable care act., Health Affairs 39(3): 502–508.

Niaksu, O. (2015). Crisp data mining methodology extension for medical domain, Baltic Journal of Modern Computing 3(2): 92.

Whatley, M. (2022). Ordinary least squares regression, Introduction to Quantitative Ana lysis for International Educators, Springer, pp. 91–1