CHAPTER THREE

3.1 RESEARCH METHODOLOGY

This study will show how to predict insurance charges using:

I. Actuarial Techniques

II. Predictive Analytics

3.1.1 ACTUARIAL TECHNIQUE

In predicting insurance premium using conventional actuarial formular, we make use of secondary data gotten from the archives of Lead way Insurance plc. The researcher also makes use of he term insurance techniques payable at the moment of death and at the end of the year of death. Which is written as:

$$\bar{A}^{i}_{x:n} = E[Z] = E[Z_t] = \int_0^n b_t v^t {}_tp_x \mu_{x+t}\, dt$$

The actuarial symbol can be defined as :

$\bar{A}^{i}_{x:n}$	=	Actuarial present value of a term insurance payable at the moment of death for a life age (x) to n-years
b_t	=	The benefit function for a life age (x)
v^t	=	The discounting function for a life age (x)
${}_tp_x$	=	The probability that a life age(x) will survive to (x + t) years
μ_{x+t}	=	The force of mortality for a life age(x) to age (x+t)

3.1.2 PREDICTIVE ANALYTICS TECHNIQUE

In predicting insurance premium using predictive analytics, we make use of secondary data gotten from https://www.kaggle.com/raghupalem/insurance-charges.

The researcher also make use of data science tools including:

1. Python computer programming language - The whole code will be written in python
2. Python libraries:
    - MATPLOTLIB: for data visualization (graphs, histogram, etc)

- NUMPY: extension module that offers fast, precompiled functions for numerical routines
- SCIPY: Python module for linear algebra, integration, optimization, statistics, and other frequently used tasks in data science
- PANDAS: contains high-level data structures and tools that are perfect for data wrangling and data munging
- SCIKIT-LEARN: This library provides a common set of machine learning algorithms through its consistent interface and helps users quickly implement popular algorithms on data sets. It also has all the standard tools for common machine learning tasks like classification, clustering, and regression.
- STATSMODELS - Python module that provides classes and functions for the estimation of many different statistical models, as well as for conducting statistical tests, and statistical data exploration.

The researcher aims at predicting insurance premium charges using the **Multiple Linear Regression** tool (contained in the python scikit-learn library). This will be implemented using software technologies listed above and each steps will be explained and visualization will be made where neccessary.

3.1.2.1 PREDICTIVE ANALYTICS TECHNIQUE MODEL : MULTIPLE LINEAR REGRESSION

When simple linear regression is extended to include multiple features or independent variables, it is called multiple linear regression.

As a predictive analysis, the multiple linear regression is used to explain the relationship between one dependent variable and two or more independent variables.

The formular for the Multiple Linear Regression is :

$y = \beta_o + \beta_1 x_1 + \beta_2 x_1 + ... + \beta_n x_n$

Where:

- y = Dependent Variable
- x = Independent Variables
- β_o = y-intercept (constant term)
- β_n = slope coefficients for each Independent variable

$\beta_1 + \beta_2 + \ldots\ldots + \beta_n$ are the model coefficients needed to predict insurance premium charges. The values of these coefficients needs to be 'learned' or trained after which they can be used to predict insurance premium charges.

The data obtained will be used to answer the research questions and hence solve the stipulated problems.

Features/independent variables of our data are:

- Age
- Sex
- Body mass index
- Children
- Smoker
- Region

Response/Dependent variable

- Premium charges

The number of rows are 1339

Using MULTIPLE LINEAR REGRESSION tool contained in Python's Scikit-Learn Module, for a new policy holder, given a set of input (ie our features/independent variables), we want to predict insurance charges (i.e our response/dependent variable)

3.1.2.2 USING THE MODEL FOR PREDICTION

Using the model for prediction, the steps to follow includes:

- Data processing
- Model evaluation
- Predicting insurance premium charge

1. DATA PROCESSING

Processing the data involves the following:

i. Import The Dataset

This involves importing our dataset and reading it into our data frame where we can view the table of data and get more information about our dataset including the labels of our columns, the number of rows and columns.

ii. Clean The Dataset

Our dataset might contain Some missing values and Categorical values.

a) Missing Values

Rows that contains missing values, might be deleted, or each missing values will be replaced with the same number.

b) Categorical Values

We need to represent all data numerically. Some of our independent variables contain categorical data eg sex, smoker and region. Each categories will be encode with a binary value (0 for male, 1 for female).

For data with more that two categories, each categories will be encoded with number 1, 2, 3 ...etc depending on the number of categories eg for region we have 4 categories and each will be replaced accordingly : 'southeast' = 0, 'northwest' = 1, 'southwest' = 2, 'northeast' = 3

iii. Visualizing Relationships

We visualize the relationship between the dependent (insurance premium charge) and the independent variables (sex, age, no of children,smoking habit, region, body mass index) using Scatterplots Package in the Python Matplotlib Library.

The relationship between insurance charges and each of the independent variables will be represented on scatterplots graph.

After visualizing the relationship between the premium charges and sex, age, no of children,smoking habit, region, body mass index, the following questions about the dataset arise:

a) Is there a relationship between:

i. Sex and charges?

ii. Age and charges?

iii. Smoker and charges?

iv. Body mass index and charges?

v. Children and charges?

vi. regions and charges?

b) If there is, how strong is that relationship?

c) What is the effect of each independent variable on the dependent variable?

d) With this dataset, can premium be predicted for a new life insurance policy holder?

e) How accurate will our prediction be?

This questions will be answered in the model evaluation stage

2. MODEL EVALUATION

At this stage, we evaluate the model with the following steps:

i. Preparing The Data For Training The Model:

At this stage we select independent variables to used in our model by evaluating their p-values.

HYPOTHESIS TESTING AND P-VALUES

Here's a conventional hypothesis test:

- Null hypothesis : There is no relationship between say smoking habit of policy holders and insurance charges i.e $\beta_5 = 0$
- Alternate hypothesis : There's a relationship between smoking habit of policy holders and insurance charges ie ie $\beta_5 \neq 0$

Here's how we can test this hypothesis:

If the confidence interval does not include the null value, the p-value will be less than 0.05 and we fail to accept the null hypothesis and accept the alternate.

Thus, a p-value less than 0.05 is one way to decide whether there is likely a relationship between the feature/independent variables and the response/dependent variable. We only keep predictor in the model if they have small p-values.

We select and use independent variables that have strong relationship with the dependent variable and drop the column or not include in our model the independent variables that have weak or no relationship with the dependent variable.

ii. Estimating or "Learning" The Model Coefficient:

Python's Statsmodels tool is used to estimate the model coefficients for all independent variables present in our model. We do this by first creating a fitted model using the Python's Statsmodels Package and from the summary of the model we can access the model coefficient for each independent variable used in our model.

iii. Interpreting Model Coefficient:

To interpret a model coefficient (say β_1), we consider the effect it has on the dependent variable/premium charge when the corresponding independent variable (say x_i) is increased by a unit.

iv. Splitting The Data Into Training and Testing Sets:

A higher percentage is used for training the model while a smaller portion of the data is used to test the model.

v. Training and Testing The Model:

The model is trained to predict the known outputs/premium charges and later tested using the test data. The test data is used to test the accuracy of the model.

vi. Model Evaluation Using R-squared:

To measure how far a set of numbers is spread out we use the variance. Variance describes how much a random variable differs from its expected value.

To evaluate the overall fit of a linear model, we analyse the R-squared. The proportion of variance in the observed data that is explained by the model is the R-squared. The R-squared is between 0 and 1 and the higher the R-square the better because it means that more variance is explained by the model. The R-squared is most useful in comparing different models.

vii. Summary Of The Fitted Model:

The summary describes the model. It reveals more information about the model. This information will be displayed on a table.

3. PREDICTING INSURANCE PREMIUM CHARGE

   We predict insurance premium charge for a new policy holder following the steps:

   I. Import the linear regrssion model from the Python's Scikit-Learn Library

   II. Instantiate or run the model

   III. Fit the data into our model

   IV. Predict for a new policy holder using model coefficient

3.2 RESTATEMENT OF THE RESEARCH QUESTIONS

The following research questions shall be answered at the end of the study:

- To what extent does insurer calculate effective premium rating?
- Does insurer consider experience rating in calculating premium for a policy holder?
- Is premium rate in the life office varying with that of the non-life?
- Which method is more efficient and effective; using conventional actuarial formulas or using predictive analytics?

3.3 METHOD OF STUDY

The method of this study is to determine the relationship that exist between premium level and the rate charged in the insurance industry considering data availabiliy

This will be carried out using two different techniques:

- Actuarial Techniques (where there is insufficient data)
- Predictive analytics (where there is sufficient data)

Data was gotten from the archives of Leadway Insurance Plc and Kaggle.com

1. Actuarial Techniques

Data from the archive of the company leadway insurance plc were collected, evaluated and finally tabulated for analysis and interpretation using the calculation of term insurance and annuities in continuous and discrete form to calculate the premium.

2. Predictive analytics

Data from kaggle.com collected, read into the computer software after which:

- Categorical data were encoded with binary values 0 and 1,
- Relationships between the dependent variable (premium charge) and the independent variables (age, sex, number of childer, smoker etc) were visualized with the aid of graphs,
- The linear regression model was imported and instantiated
- We fit in our dependent and independent variables in order 'learn' our model coefficients
- We predict insurance premium charge using our model coefficient

3.4 CHARACTERISTICS OF THE STUDY POPULATION

During our investigation, we shall be considering the following:

- Using Actuarial Techniques
  - The insured who took up the policy between a certain periods of time.
  - The entire insured sum assured, claims and the interval between the start of the policy and date of maturity claims will be considered
- Using Predictive Analytics
  - The age, if policy holder is a smoker or not , sex, region, body mass index, number of children and their relationship with premium charge will be considered.

3.5 DATA COLLECTION METHODS AND SOURCES

Data was collected from an Insurance company and Kaggle.com

The data collected from the insurance company will be processed to determine or predict premium using Actuarial Techniques.

The data collected from an online data repository; kaggle.com will be processed predict premium charge using Multiple Linear Regression / Predictive Analytics

3.6 SAMPLING PROCEDURE

In predicting insurance premium using **Actuarial Techniques,** secondary data was used. From the original data of 31 policy holders of Leadway insurance company which gives their policy number, sum assured, maturity proceeds. The sample used is the entire population of 31 people.

In predicting insurance charges using **Predictive Analytics,** secondary data was used. From a data set gotten from kaggle.com which contains 1339 rows of data which gives sex, age, body mass index, smoker, region, premium charge. The sample used is the entire populaion of 1339 people.

3.7 PROCEDURES FOR PROCESSING AND ANALYSING COLLECTED DATA

For the purpose of this study the following techniques will be used:

- The Actuarial Techniques (using conventional actuarial formulas):

  The actuarial present value of a term insurance payable at the moment of death and at the end of the year of death and the actuarial present value of annuity will be used to analyze data to estimate the benefit premiuim of each policy holder at the moment of death and at the end of the year of death

- The Data Science Technique (using predictive analytics)

  The multiple linear regression model will be used. The model coefficient will be determined using a statistical tool contained in the computer software programming lanuage- Python. This model coefficient will be used to predict premium charge.