INSURANCE PREMIUM PREDICTION

- BY:
- BARNALIKKA PRADHAN
- PUNITH B C

Objective:

To create a insurance premium predictive model to estimate the cost of insurance for an individual based on the user inputs given by them on the following 6 attributes of age, sex, body mass index, no of children, smoker or non smoker, region they belong to.

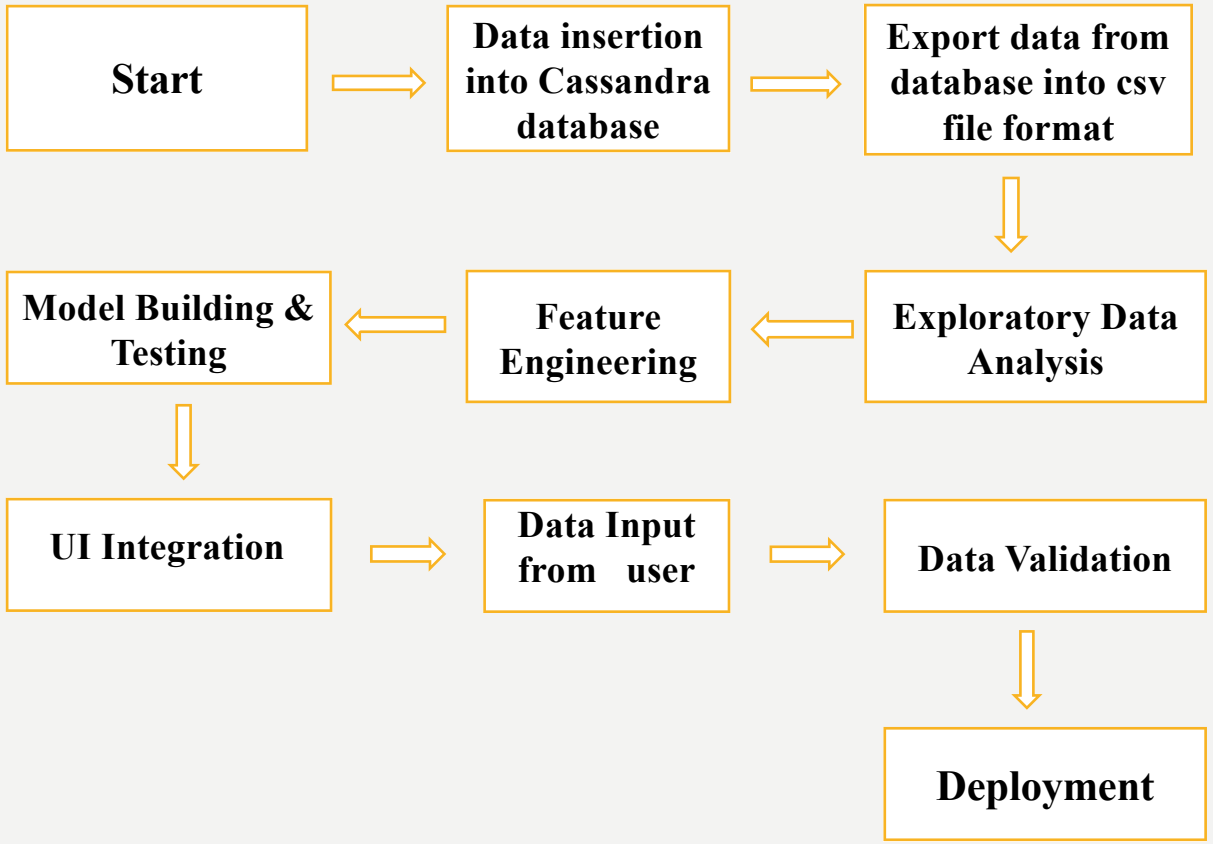
Benefits:

- Gives an estimated cost of insurance premium pertaining to their individual physiological features.
- Help the individual plan their insurance premium for a given period of time.
- Gives them the clarity about the best suited insurance plans for them.
- To learn about the scope of benefits included in it, the plans cost sharing requirements and the extent of coverage.

Data Sharing Agreement：

- Sample file name (dataset.csv)
- Length of date stamp(8 digits)
- Length of time stamp(6 digits)
- Number of Columns
- Column names
- Column data type

Architecture Design



Data insertion into Cassandra Database:

- First you connect to the Cassandra database using security connect bundle along with the given client id and client secret .
- Then you create the database with the personal keyspace you created.
- Secondly, table is created named “experim” in the cassandra database.
- Then you insert the dataset named insurance.csv into the table called ‘experim’ in cassandra database.

Model Training:

Export Data from Cassandra database:
The accumulated data from database is exported in csv file format for model training.

Exploratory Data Analysis:
One of the best ways to see the relation among the variables is by visualization.

- In the univariate analysis we have analyzed each of the variables using distplot to see whether expenses has a normal distribution, countplot to see age of insurers, no of children, region to which they belong ,category of BMI and piecharts to check the no of smokers to non-smoker and male to female.

- For bi-variate analysis we have tried to see the relationship between expenses to no of children these individulas have and to region they belong to.

- For multi-variate analysis, we try to explore the relation of whether the person is a smoker or not to his/her expenses related to age for one scatterplot and the other to the bmi

Feature Engineering:

-We have encoded the categorical variables into numeric namely:

- sex – male:1 and female :0,
- smoker – no:0 and yes:1
- region the individual belong to - 'southwest': 0, 'southeast': 1, 'northwest': 2, 'northeast': 3

-After encoding the train test split, standardised scaling has been done to scale down all the features relatively on a similar scale to a zero mean and standard deviation of one (unit variance) which make it easier for the ML to function.

Model Selection:

Coming to the part of model building, three machine learning algorithms, i.e Decision Tree Regressor, Random Forest Regressor, Gradient Boost Regressor will be fine-tuned using Grid search CV and the one with the highest accuracy score which in this case is the Random Forest classifier with a test accuracy of 84 per cent will be used for predicting the cost of insurance premium.

Q & A:

Q1) Explain about the project?

In this project we have created a web application to predict the cost of insurance premium for individuals taking their physiological features as inputs mainly their sex, age,bmi, no of children, region they belong to and whether they have the habit of smoking or not.
Prediction of the cost will give the individuals the clarity about the best suited insurance plans for them for a certain period of time under which they can look at the scope of benefits included, the plans cost sharing requirements and the extent of coverage.

Q2) What's the source of data?

The dataset is taken from kaggle problem statement.

Q3) What was the datatype?

The data was a combination of categorical and numerical values.

Q4) What's the complete flow you followed in this Project?

Refer 5th slide for a better understanding.

Q5) How logs are managed?

We have used logs for database insertion, model training, model selection and prediction log.

Q6) What steps were used for data pre-processing/feature engineering?

- Converting categorical to numeric variables.
- Standard scaling of the data to bring it to a fixed range.

Q7) How training was done and what models were used?

Before performing scaling operation,, the dataset was already divided into train and test data after which algorithms like Decision Tree Regressor, Random Forest Regressor, Gradient Boost Regressor which were fine-tuned using Grid search CV were used to train and test the model and the one with the highest accuracy score was choosen for prediction of insurance premium.

Q8) Which Tool You Are Used For Implementation This Model?

1) Ide : Pycharm
2) Cloud : AWS
3) Data Base : Cassandra

Q9) What are the different stages of deployment?

- Firstly you login into the AWS management console, then you choose Elastic Beanstalk, which is a service provided by AWS for deploying and scaling web application.
- In the Elastic beanstalk you create a new environment, select environment tier as web server environment and give the application name & python version used.
- Meanwhile you open Code pipeline in AWS, and you click on create a pipeline, give a pipeline name, connect to GitHub version 2 as your source provider thus creating a connection giving name of your GitHub repository.
- Then you choose your deployment provider as AWS Elastic Beanstalk, give your application and environment name.
- Once these steps have been completed, the source i.e GitHub connects to deploy your model in elastic beanstalk.
- Once the deployment has succeeded, then when you click on the environment name you will get the link to the web application.