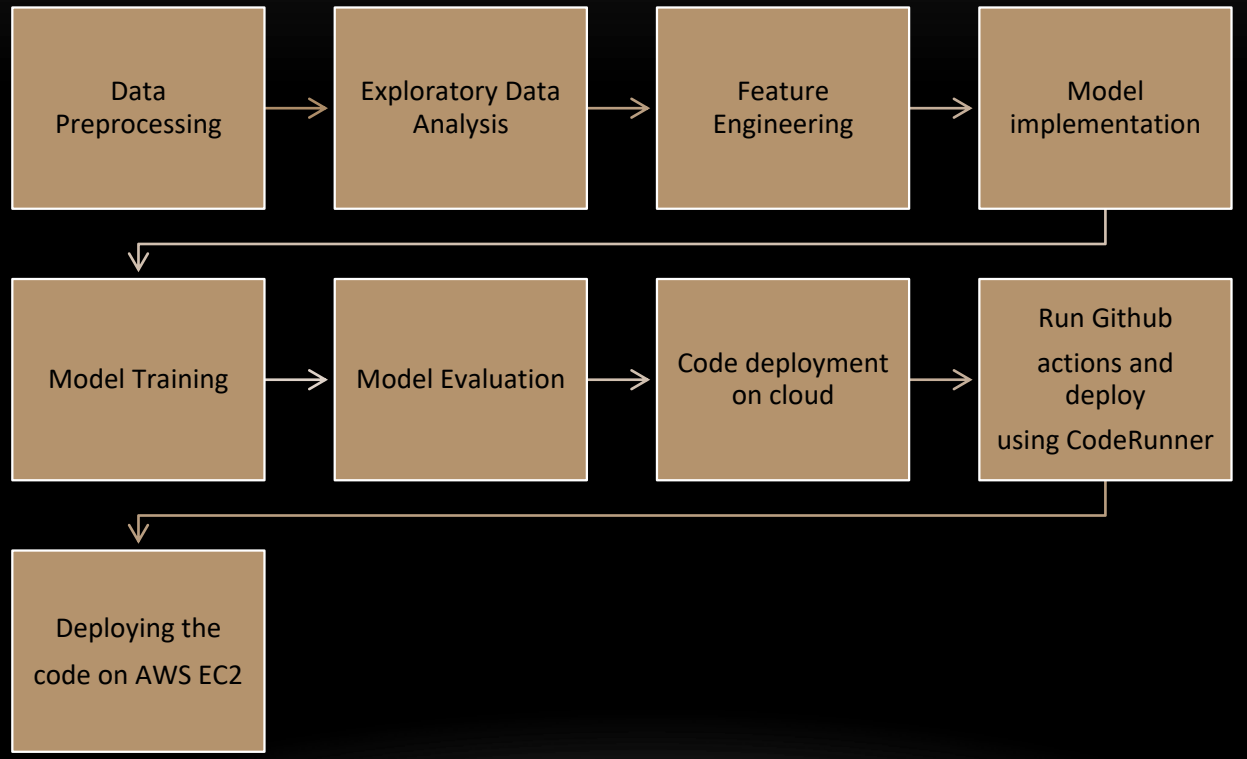
Insurance Premium Prediction…

- **Objective:**
- Development of a predictive Regressor model to find premium of insurance one has to pay depending on conditions mentioned. This model requires certain set of information like Age of the Person, Gender, Number of Children, Region where the person resides, whether the person is a Smoker or Not and the person's Body Mass Index (BMI).

- **Benefits:**
- The model predicts the premium the user has to shell to secure their family against medical expenses.

ARCHITECTURE



Data Validation and Data Transformation :

- Name Validation – Validation of the file using the pandas read.
- Number of Columns – Validation of number of columns present in the .csv files, and if it doesn't meet the specification then the file is removed.
- Name of Columns - The name of the columns is validated and should be the same as given in the schema file. If not, file to discarded.
- Data type of columns - The data type of columns is given in the schema file. It is validated when we insert the files into Database. If the datatype is wrong, then the file is moved to recycle
- Null values in columns - If any of the columns in a file have all the values as NULL or missing, we discard such a file discarded…

Model Training:

- Data Import from local memory :

  The accumulated data from memory is exported in csv format for model training

- Data Preprocessing
  - Performing EDA to get insight of data like removing the urls, html_tags, word corrections,
  - Removing stop words, tokenization, stemming the text data and pad the sequences with max length of text.
  - Check for null values in the columns. If present impute the drop them.
  - Encode the categorical values with numeric values.

Model Selection:

- Deploying as many number of models as possible like

  - Linear Reg, Random Forest, Decision Tree, Adaboost, Grad-boost.

  - Use GridsearchCV to hyper tune the models.

  - We find the best model by comparing the accuracy scores and classification reports.

  - By the final model, train and evaluate the model using classification report or accuracy score or confusion matrix.

Prediction:

- The testing files are shared in the batches and we perform the same Validation operations ,data transformation and data insertion on them.
- The accumulated data from local file system is exported in csv format for prediction
- We perform data pre-processing techniques on it.
- Best model classifier model created during training is loaded for the preprocessed data is predicted
- Once the prediction is done for all the testing data. The predictions are saved in csv format and shared.

Data Insertion in Database:

- Table creation :- Table name “predictions" is created in the database for inserting the files. If the table is already present then new files are inserted in the same table.
- Insertion of files in the table - All the files in the user interface are to be inserted along with the prediction result in the above-created table.

Q & A:

Q1) What’s the source of data?

The data for training is provided by the client in csv format

Q 2) What was the type of data?

The data was the combination of Numerical and Categorical values.

Q 3) How logs are managed?

We are using different logs as per the steps that we follow in validation and modeling like File validation log , Data Insertion ,Model Training log , prediction log etc.

Q 4) What techniques were you using for data pre-processing?

- Dropping unwanted columns
- Visualizing relation of independent variables with each other and output variables
- Converting categorical data into one hot representation values.

Q 5) How training was done or what models were used?

- Divide the data to training and validation data.
- Algorithms like Linear Reg, Random Forest, Decision Tree, Adaboost, Grad-boost from these find and save final model .

Q 6) How Prediction was done?

The testing files are shared by the user .We Perform the same life cycle till the data preprocessing. In the end we get the accumulated data of predictions.

- Q 7) What are the different stages of deployment?
-

The code was first committed on Github. The pipeline was created between Git and Code Runner. Then the code was deployed to the AWS. This process is been done using Github Actions. Hence Continuous Integration, Continuous Delivery and Continuous Deployment.

THANK YOU…