Insurance Premium Prediction

Sahil & Siddhi

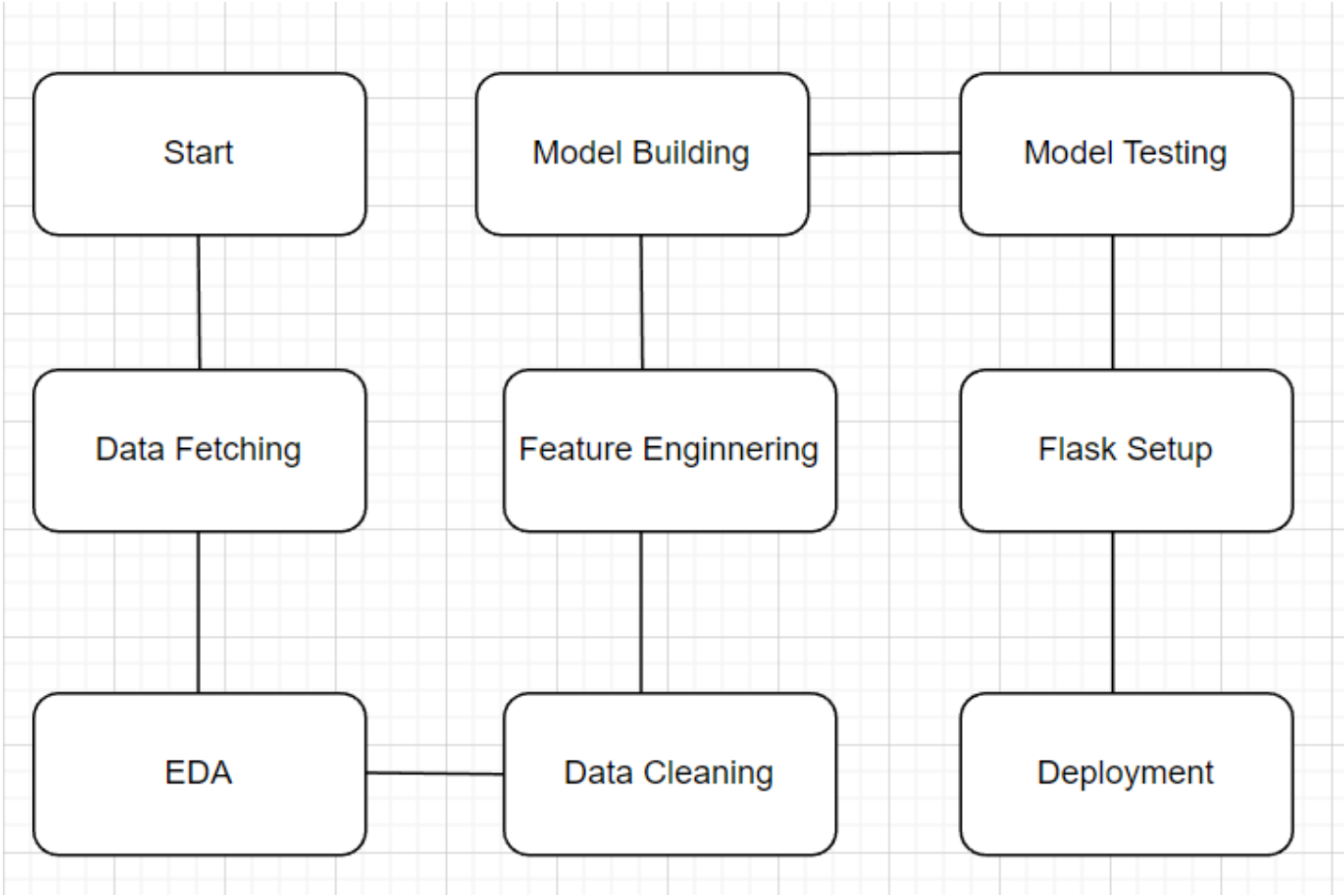
Objective：

The goal of this project is to give an estimate of how much they need on their individual health situation and Build a solution that should able to predict the premium of the personal for health insurance.

Benefits :

- Gets idea about how much amount required annually according to their own of health status.
- This can help a person in focusing more on the health aspect of an insurance.
- Help in giving premium of health insurance.

Architecture



Data Collection and validation

- The dataset was taken from the Kaggle competition page.
- Data type of columns – Validating the data type of the columns if wrong, then it was corrected.
- Null values in columns – Validating the column in the dataset have null values or missing information.

Model Training

➢ Data Pre-processing:

- Performing EDA to get insights of the data like identifying distribution, outliers etc.
- Check any null values present in the dataset. If present then impute those null values.
- Encode the categorical features/columns.
- Perform Standard Scalar to scale down values.

Model Selection

After pre-processing and model training, we find the best model for premium prediction. The model is trained on multiple regression algorithms like Linear Regression, Decision Trees, Random Forest, Gradient Boosting, Extreme Gradient Boosting and K-Nearest Neighbors (KNN). After prediction we will find accuracy of those predictions using evaluation metrics like RMSE (Root mean squared error) and r2_score (R-squared).

Predictions

Then all the trained models were used for validating test set.

We perform pre-processing techniques on it.

The best RMSE and r2 score model were saved for developing API for prediction of premium.

Q & A

Q1) What is the source data?

The source of the data is Kaggle. The data is in the form of ‘csv’ file.

Q2) What was the type of the data?

The data was combination of categorical and numerical values.

Q3) What’s the complete flow you followed in this project?

Refer the 3rd slide for better understanding

Q4) What techniques were you using for data pre-processing?

Visualizing relation of independent variables with each other and dependent variable.

Checking distribution of Continuous variables.

Checking any null values present in the dataset.

Converting categorical data into numeric values.

Scaling the data.

Q5) How training was done or what models were used?

Before training the model the dataset is divided into training set and testing/validation set.

The scaling was performed of training and validation set.

The categorical columns were converted into numeric values.

Algorithms like Linear Regression, Decision Trees, Random Forest, Gradient Boosting, KNN, and Extreme Gradient Boosting were used for model training and based on RMSE & r2_score the Gradient boosting model is saved for Validation.

Q6) How prediction was done?

On the basis of trained model, the prediction was performed. We also created API interface for estimating cost of premium on the basis of personal health information/status.

Q7) What are the different stages of deployment?

When the model is ready we deploy it in Heroku platform.