ADULT CENSUS INCOME PREDICTION

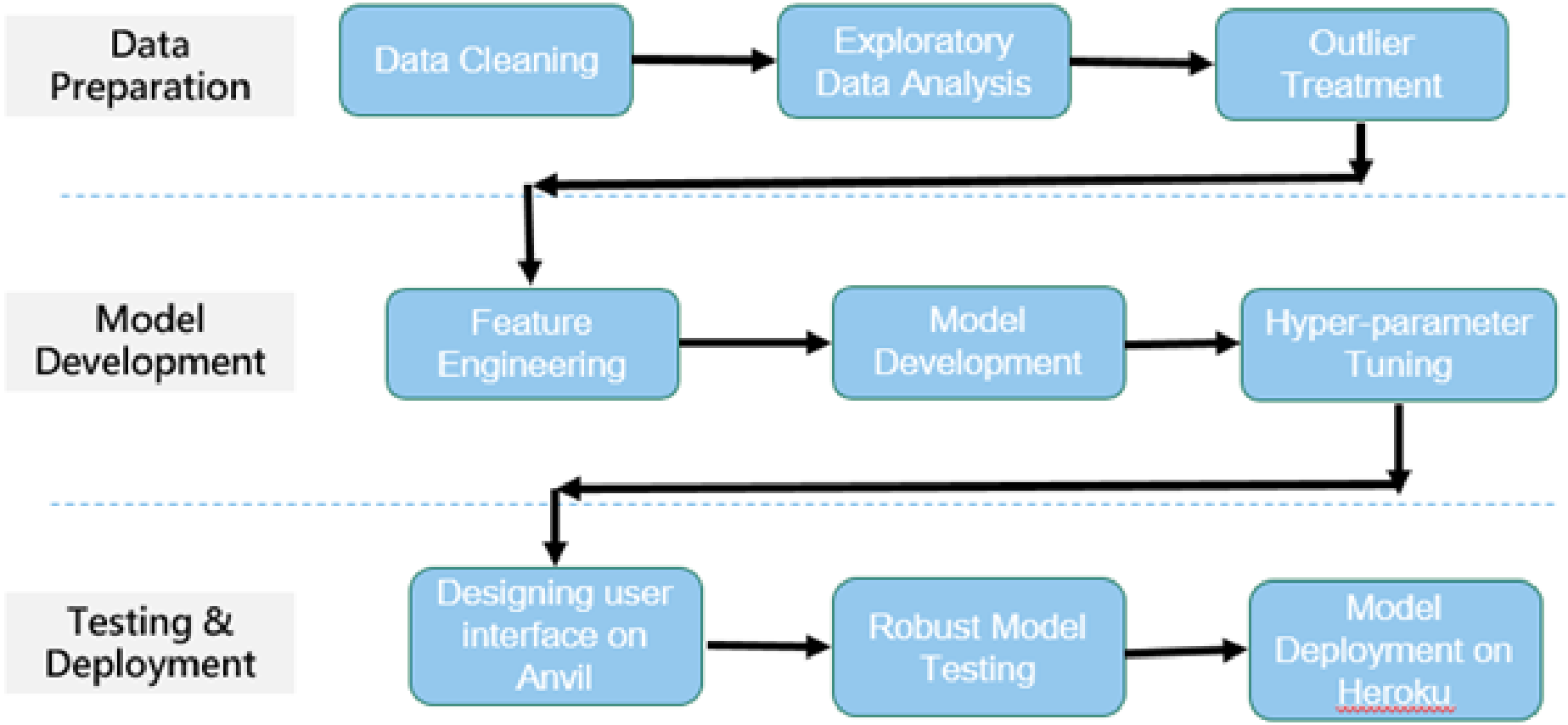
Objective:

Development of a predictive model to classifiy whether a person earns more than 50K dollars per annum or not.. The model will determine the same using some parameters like age, workclass, educational qualifications, country, etc

Benefits:

- Can be used by Governmental / Non-Governmental / Private agencies where people needed to be classified on the basis of their annual income
- For eg : in granting scholarships or waivers to needy students, for different schemes by government for poor ones, etc

Architecture



Data Validation and Data Transformation :

- Missing Values – All the missing values were replaced with the value being repeated the most number of times.
- Numerical Columns - All the numerical features were standardized, preventing any data leakage by using data pipelines.
- Categorical Columns - Either label encoding or one hot encoding was done to treat the categorical features

Data Insertion in Database:

- Cassandra Database - The dataset was imported to cassandra database from where we can access it with the help of python
- Insertion of files in the table - All the data is uploaded into a table named “adult_data” into which is present inside a database named “my_database”.

Model Training:

- Data Export from Db :

  The accumulated data from db is exported to python and read using pandas

  - Performing EDA to get insight of data like identifying distribution , outliers ,trend among data etc.
  - Check for null values in the columns. If present impute the null values.
  - Encode the categorical values with numeric values.
  - Perform Standard Scalar to scale down the values.

- Model Selection

  Different classification models were compared and hyperparamater tuning was done via gridsearchcv on the best performing one that is CatBoost Classifier

- Prediction

  The model is made in such a way to maximise the accuracy and also other performance metrics so that the predictions are as accurate as possible

  The average accuracy after cross validation was observed to be 84 percent and average f1 score as 69 in 10 validations.

Q & A:

Q1) What's the source of data?

The data for training is provided by the client in the form answers to certain questions asked which the user has to input.

Q 2) What was the type of data?

The data was the combination of numerical and Categorical values.

Q 3) What's the complete flow you followed in this Project?

Refer slide 3 for better Understanding

Q 5) How logs are managed?

Following are the logs that we are using :

modeling like, Data Insertion log, Model Fitting log, prediction log, etc.

Q 6) What techniques were you using for data pre-processing?

- Removing unwanted attributes
- Visualizing relation of independent variables with each other and output variables
- Checking and changing Distribution of continuous values
- Removing outliers
- Cleaning data and imputing if null values are present.
- Converting categorical data into numeric values.
- Scaling the data

Q 7) How training was done or what models were used?

- First, we started with data cleaning, EDA and feature engineering
- Then, outliers and ambiguities were removed from the data and categorical features data transformation was applied for categorical columns like one hot encoding, label encoding, etc
- Data pipeline was created to implement data scaling, upsampling using SMOTETomek and an estimator to prevent any data leakage
- Catboost model was used as the best estimator which was then used for production followed by hyperparameter tuning

Q 8) How Prediction was done?

Some questions were asked to the client like his age, qualifications, etc and his responses are taken as inputs which are then feeded to the model as a single test case and the predictions are then returned on the clients screen after a interval of three seconds in which the data pipeline processes the input data to get the output