***ML Income Prediction***

1. **Problem Statement**

* To build classification model to predict which salary/ income category a particular individual belongs based on personal information.
* Companies can use this type of classification to sell/ promote products based on the salary range of the particular individual. It can help to proper allocation of resources and better targeted based marketing

1. **Description of data**

* Dataset contains continuous values and categorical values.
* Columns inside the dataset are:

1. age (continuous): age of individual person
2. workclass (categorical): private, government, etc. sector of employment
3. fnlwgt (continuous):
4. education (categorical): Type of degree
5. education-num (continuous): Number assigned to the type of degree
6. marital status (categorical): married, divorced, never married
7. occupation (categorical): type of industry eg. Tech support, sales
8. relationship (categorical): husband wife, son, not in family
9. race (categorical): white, black, asian
10. sex (categorical): male, female
11. capital-gain (continuous): amount gained
12. capital-loss (continuous): amount lost
13. hours-per-week (continuous): works per week
14. native-country (categorical): country belongs
15. Income (Target Label): <=50k or >=50k
16. **Application architecture and module division**

* Bigger problem: How to classify the data to decide which class the person belongs to?
* Better to break down the development into small parts so that changes to be made in end of module doesn’t affect the other modules. Multiple members part of the project so better to divide the modules
* Broker into small subgroups:

1. How to read the data
2. How to validate the data
3. How to do data preprocessing and how to train a model on the data
4. How to do hyperparameter tuning for the model

* **Step 1**: Data ingestion

1. Data for training - client provides or stores the data needed at a particular location, aggregate multiple data sources
2. Data validation – discuss with client the datatype of variables, number of variables, whether any columns contain only null values
3. Data transformation – missing values conversion to null, categorical values in commas or “” and maybe not accepted in DB
4. Data insertion in DB – after transformation insert the data inside the database for further development

* **Step 2**: Training Pipeline / Step

1. Export the data in CSV from DB and csv acts as train data
2. Data preprocessing – perform EDA, check if there are any null values present, convert categorical values into numerical values, if data is imbalanced or normalized
3. Data clustering - to increase the accuracy of the model we divide the data into individual clusters and build model for each cluster separately
4. Hyperparameter tuning - to increase the performance of the individual model selected for each cluster
5. Model saving – save the model for each cluster individually

* **Step 3:** Deploy on cloud

1. Create metadata for pushing the app onto the cloud server
2. Start and test the application

* **Step 4**: Prediction Pipeline

1. Data validation – discuss with client the datatype of variables, number of variables, whether any columns contain only null values
2. Data transformation – missing values conversion to null, categorical values in commas or “” and maybe not accepted in DB
3. Data insertion in DB – after transformation insert the data inside the database for further development
4. Export the data in CSV from DB and csv acts as train data
5. Data preprocessing – perform EDA, check if there are any null values present, convert categorical values into numerical values, if data is imbalanced or normalized
6. Data clustering - to increase the accuracy of the model we divide the data into individual clusters and build model for each cluster separately
7. Call the model for specific cluster number stored
8. Make prediction and export the prediction in a csv file

* **Step 5:** Model retraining

1. When new patterns detected these changes must be aggregated to the model.
2. Provide the prediction + train data to the model for retraining
3. Logging and monitoring framework
4. **Code**:

Main .py -🡪 1. Validation step – read data, validation, transformation, insert into DB, export

to csv file.

* 2. Training – read train data, data preprocessing, data clustering, model finding,

Model tuning, deployment

* 3. Prediction – validation,

Prediction – model saved loaded into memory and make predictions

* Data for training: User provides different training batch files
* Synchronize logging, asynchronous – code doesn’t wait for completing the logging faster and individual
* Data Validation: Whether the data sent by the client is valid or not as per requirement given. File name is correct or not based on agreement, if we reject the data push into bad folder else put in good data folder. We use the schema files created. We delete the good and bad directory as the good data is stored inside the database. Check the length of the timestamp given in the file name format, check the number of columns present in the data, check if any column contains just all column values as null values.
* Data Transformation: DB doesn’t accept Nan value so we transform to NULL using fillna. Add quotes to all values present in the categorical values for insertion into the DB.
* Data Insertion in DB: Put the good raw data inside the DB using the datatypes given inside the schema. Export the good raw data table into csv.
* Data preprocessing: Use the exported csv file as input. Perform EDA to understand the data and what processing is needed for the data.

Strip or remove the white spaces in categorical values.

Replace the “?” with NaN in the dataset and check number of missing values.

Impute the categorical columns with missing values using categorical imputer (same like mode). Drop education as the education number is already present.

Separate categorical data and numerical datatypes.

Encode the target income column.

Perform dummy encoding for the categorical variables.

Perform standard scaler for numerical datatype.

Separate into Feature and target and check the balance of the data. Balance the data using randomoversampler.

* Model Training: Fetch the data and perform the preprocessing decided after doing the EDA. Separate the target and features. Perform clustering on the cleaned data and create clusters. We can apply cluster specific algorithms for better performance. Loop individual clusters and apply different models. Perform standard scaler on the input data.

Logistic regression, DT, RF, SVM, NB, KNN, XGBoost algorithms can be used for classification.

* We perform hyperparameter tuning for the selected algorithms and check the performance of the model. Then save the model for further prediction. We used naive bayes and XGBoost. If same label classes present then we use the accuracy score else we can use the roc\_auc\_score for the check. Save the model for each cluster created.
* Prediction –

1. perform validation – for filename, no of columns, all of null values
2. perform transformation – replace nan, insert into db, export csv as input
3. data preprocessing – missing values and imputation, std == 0 dropped, drop unnecessary columns
4. Perform clustering to determine which cluster it belongs to using Kmeans.predict
5. Then based on cluster assigned use the respective model for each individual cluster. Reassign the encode class category back on the predicted file.

* Index.html – default for every browser it call webage it returns httpget

Render template to display respective html pages.

* Deployment for cloud 🡪

1. Requirements.txt – import of packages are included here, as cloud needs instruction for cloud deployment (pip freeze requirements.txt)
2. Go to GCloud console and create a new app
3. Goto IAM and admin 🡪 manage resource 🡪 create a project

(Study Material) Goto App engine 🡪 dashboard 🡪 choose correct project 🡪 choose start tutorial 🡪 choose language 🡪 click start 🡪 clone from git and deploy

Use download gcloud cli 🡪 open cmd 🡪 gcloud (run this command, not running, go to environment variables and add gcloud path)

1. Change directory and goto the project directory location 🡪 gcloud init 🡪 login 🡪 enter choice (2) 🡪 choose the project id 🡪 gcloud app deploy app.yml –project name 🡪 select the region 🡪 enable the access