ADULT CENSUS INCOME PREDICTION

Detailed Project Report

Saakar Sengar
Shreyansh Jain

OBJECTIVE

- The main aim of this project is to The Adult Census Income Prediction web application to classified the income category of either greater than 50K Dollars or less equal to 50K Dollar category of the person by using classification based Supervised Machine Learning algorithms. This Website helps the user predict whether a person has an income of more than 50K a year or not. The purpose of this application is to discover patterns in income data and then make predictions based on of ten complex patterns to detect and analyse trends and help solve problems.
- For this objective Machine learning is effectively a method of data analysis that works by automating the process of building data models.

Benefits

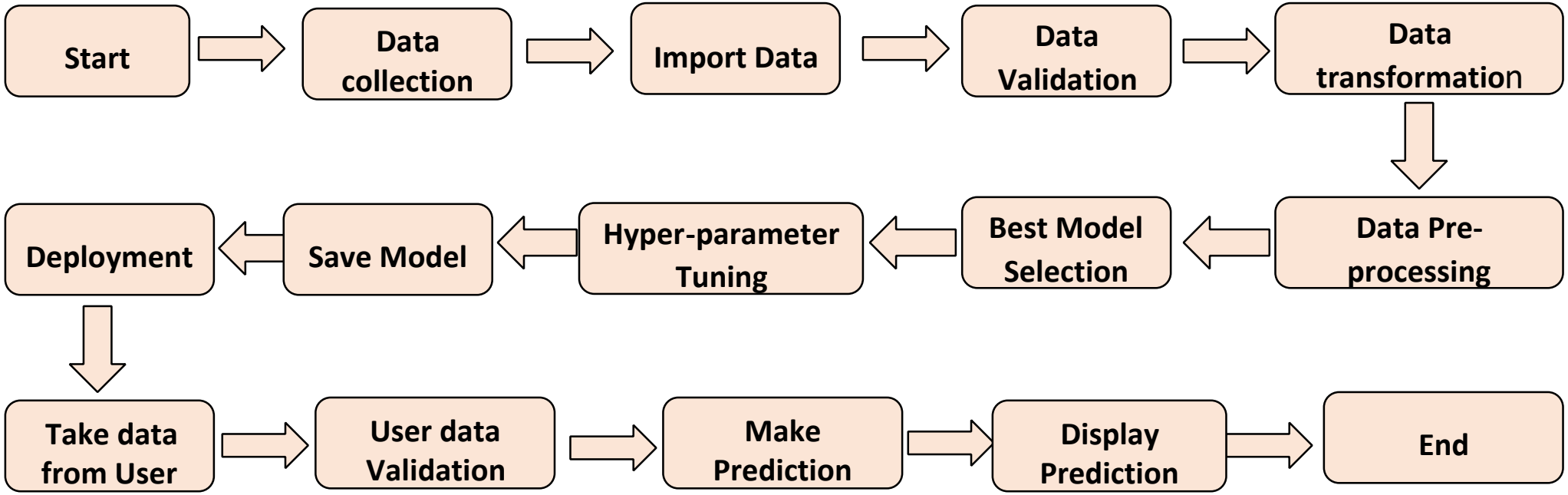
- The primary purpose of machine learning based Adult Census Income Prediction web Application to discover patterns in the user Income data and then make predictions based on these and intricate patterns

- Enables complex and larger data to be processed and analysed along with the desired results being achieved such as determining Income trends.

- Make use of the ever increasing amounts of Data being gathered by manipulating and analysing it without heavy human input.

- Now can help identify ways for income prediction to be made government so improving efficiency and maximizing the income of a person. Get idea of income of person either greater than 50K Dollars or less equal to 50K Dollars.

ARCHITECTURE



DATA Collection, Import and Validation –

- **Data Collection** – as per data is provided by ineuron.ai according to the problem statement, so first we downloaded it on our local pc in the form of .csv file from internship portal and save it in our workspace.
- **Data Import** – for further process we create a python module with the help of pandas library to import the data on our workspace.
- **Data Validation** –
  - Number of Columns – Validation of number of columns present in the files,
  - Name of Columns - The name of the columns is validated and should be the same as given in the schema file.
  - Data type of columns - The data type of columns is given in the schema file.
  - Null values in columns - If any of the columns in a file have all the values as NULL or missing.

Data Transformation and Pre-processing –

- **Data Transformation** –
  - **Extra Spaces** – some extra spaces are present in the some of the object (string) kind entities of the data frame which can hamper the work while doing pre-processing or time of training. So first we remove these spaces.
  - **Replacing unknown entities** – a question mark (“ ?”) kind entities are present in the 3 independent Categorical columns of data-frame, so we convert it into missing value (NAN).
  - **Separated Dependent and Independent Variables** – Separated all independent variables from the dependent Variable.

- **Data Pre-processing** –
  - **Missing Values Imputation** – all the missing values are present in categorical column, so we replaced them with MODE values.
  - **Categorical Variables handling** – all the categorical variables are transformed by using One-Hot Encoding.
  - **Outlier handling** – outlier computation and handling in the Numerical variables, those need it.
  - **Imbalanced Dataset** – Dataset is highly Imbalanced, so first balanced it by using Random Over Sampler.
  - **Feature Scaling** – all numeric Variables are on different scale, so we scaled down all the features on same scale with the help of Standers Scaler.

Best Model Selection and Hyper-parameter Tuning—

- After the Pre-processing is completed, we go for the Model training approach to find the best model for our dataset use the algorithms “Logistic Regression” , “Random Forest Classifier”, “Gaussian-NB” and "XGBoost Classifier" For every model tunned algorithms are used. We calculate the Accuracy-score for all the models and select the model with the best score.
- Similarly, the best model is selected is "XGBoost Classifier" and after hyper-parameter tuning by selecting best Parameters with the help of Randomized Search CV the model is saved in workspace for use in prediction.

Deployment –

- After Training and model Selection we go for the Deployment of the project for this we create a web application by using "**Flask** Web Application framework" for the backend development.

- **HTML** and **CSS** are used for the Frontend development.

- Flask is used for the backend, but it makes use of a templating language called Jinja2 which is used to create HTML markup formats that are returned to the user via an HTTP request. More on that in a bit.

- After creation of web application **"Heroku"** is used as a platform as a service (PaaS) that build, run, and operate applications entirely in the cloud.

PREDICTION –

- While Deployment we created a app which is using the input from user as a HTML form having fields —

Application Link: https://adult-census-income.herokuapp.com/

Q1) What's the source of data?

The data for training is taken from the internship portal of ineuron.ai.

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 4th for better Understanding

Q 4) How logs are managed?

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

Contact Information

Developed By –

- Saakar Sengar
- Shreyansh Jain