High Level Document

Adult Income Census Prediction Application

Written By	**Mohammad Sohail Parvez**
Document Version	0.1
Last Revised Date	

Content

Abstract

For this assignment, we examine the Census Income dataset available at the UC Irvine Machine Learning Repository. We aim to predict whether an individual's income will be greater than $50,000 per year based on several attributes from the census data.

1. Introduction

1.1 Why this High-Level Design Document?

The purpose of this High-Level Design (HLD) Document is to add the important details about this project. Through this HLD Document, I'm going to describe every small and big thing about this project.

2. General Description

2.1 Product Perspective

The adult income census prediction using classification based Machine Learning algorithms.

2.2 Problem Statement

The goal is to predict whether a person has an income of more than 50K a year or not. This is basically a binary classification problem where a person is classified into the >50K group or <=50K group.

2.3 Proposed Solution

The solution here is a classification based Machine Learning model. It can be implemented by different classification algorithms (Logistic Regression, Random Forest, Gradient Boosting, Decision Tree, SVM, XGBoost, Catboost, KNN and so on). Here First we are performing a Data preprocessing step, in which label encoding, feature engineering, feature importance steps are performed and then we are going to build a model.

2.4 Technical Requirements

In this project the requirements to get income classify various platforms. For that, in this project we are going to use different technologies. Here are some requirements for this project.

- Model should be exposed through API or User Interface, so that anyone can test model.
- Model should be deployed on cloud(Heroku).
- Cassandra database should be integrated in this project for any kind of user.

2.5 Data Requirements

The dataset is downloaded from Kaggle. The dataset contains 15 columns and 32000+ rows.

● **age**: the age of an individual
  ○ Integer greater than 0

● **workclass**: a general term to represent the employment status of an individual
  ○ Private, Self Emp Not Inc, Selfempinc, Federalgov, Localgov, State Gov, Withoutpay, Neverworked.

● **fnlwgt**: final weight. In other words, this is the number of people the census believes the entry represents..
  ○ Integer greater than 0

● **education**: the highest level of education achieved by an individual.
  ○ Bachelors, Somecollege, 11th, HSgrad, Profschool, Assocacdm, Assocvoc,9th, 7th8th, 12th, Masters, 1st4th, 10th, Doctorate, 5th6th, Preschool.

● **education-num**: the highest level of education achieved in numerical form.
  ○ Integer greater than 0

● **marital-status**: marital status of an individual. Married civ spouse corresponds to a civilian spouse while Married AF spouse is a spouse in the Armed Forces.
  ○ Married civ spouse, Divorced, Nevermarried, Separated, Widowed, Married spouse absent, MarriedAFspouse.

● **occupation**: the general type of occupation of an individual
  ○ Techsupport, Craftrepair, Otherservice, Sales, Execmanagerial, Profspecialty, Handlerscleaners, Machineopinspct, Admclerical, Farmingfishing, Transportmoving, Privhouseserv, Protectiveserv, ArmedForces.

● **relationship**: represents what this individual is relative to others. For example an individual could be a Husband. Each entry only has one relationship attribute

and is somewhat redundant with marital status. We might not make use of this attribute at all
    - Wife, Ownchild, Husband, Notinfamily, Otherrelative, Unmarried.
- **race**: Descriptions of an individual's race
    - White, AsianPacIslander, AmerIndianEskimo, Other, Black.
- **sex**: the biological sex of the individual
    - Male, Female
- **capital-gain**: capital gains for an individual
    - Integer greater than or equal to 0
- **capital-loss**: capital loss for an individual
    - Integer greater than or equal to 0
- **hours-per-week**: the hours an individual has reported to work per week
    - continuous.
- **country**: country of origin for an individual
    - UnitedStates, Cambodia, England, PuertoRico, Canada, Germany, OutlyingUS(GuamUSVIetc), India, Japan, Greece, South, China, Cuba, Iran,Honduras, Philippines, Italy, Poland, Jamaica, Vietnam, Mexico, Portugal,Ireland, France, DominicanRepublic, Laos, Ecuador, Taiwan, Haiti, Columbia,Hungary, Guatemala, Nicaragua, Scotland, Thailand, Yugoslavia, ElSalvador,Trinadad&Tobago, Peru, Hong, HolandNetherlands.
- **Salary** : whether or not an individual makes more than $50,000 annually.
    - <=50k, >50k

2.6 Tools Used

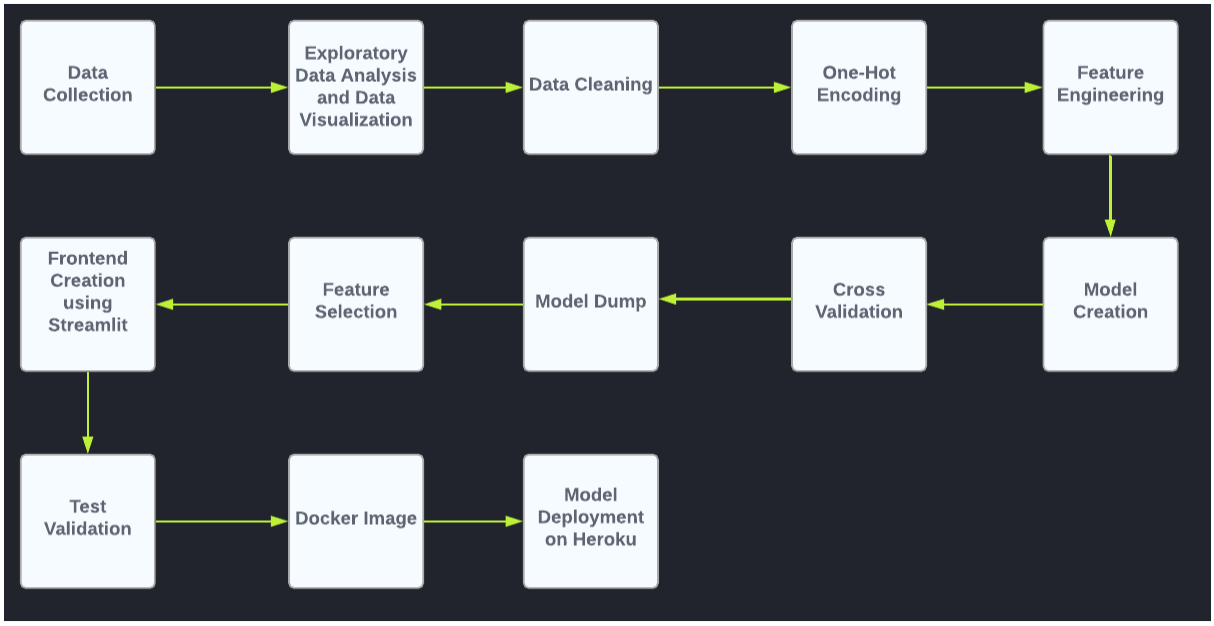
- Vscode is used as an IDE.
- For visualization of the plots, Matplotlib, Seaborn are used
- Heroku is used for deployment of the model.
- Cassandra is used to retrieve, insert, delete and update the database.
- For application Streamlit is used.
- For eda and numerical computation Pandas and Numpy are used respectively.
- Scikit-learn is used for importing different classification algorithms.

2.7 Constraints

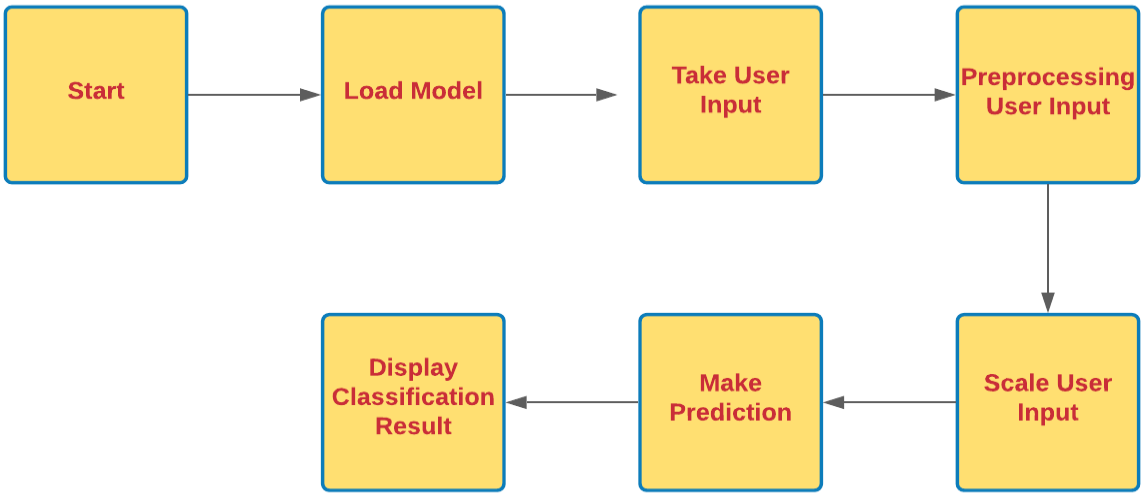
The adult income census prediction system must be user friendly, errors free and users should not be required to know any of the back-end working.

3 Design Details

3.1 Process Flow



3.2 Deployment Process



3.3 Event Log

In this project we are logging every process so that the user will know what process is running internally.

Step-By-Step Description:

- In this project we defined logging for every function.
- By logging every function in the Exploratory Data Analysis.
- Then logs all the data preprocessing functions and code.
- Logs every model algorithm to our data.

3.4 Error Handling

The project is designed in such a way that, at any step if error occurs then our application should not terminate rather it should catch the error and display that error with proper explanation as to what went wrong during process flow.

4. Performance

Solution of Adult Income Census Prediction is used to classify in advance, so it should be as accurate a s possible so that it should give as much as possible accurate classification.

That's why before building this model we followed the complete process of Machine Learning. Here are summary of complete process:

1. First we cleaned our dataset properly by removing all null value and duplicate values present in the dataset.
2. Then we performed data preprocessing steps like cleaning the data, handling the categorical values and other etc.
3. We have created a new feature column based on the columns present in the dataset before using Feature Engineering. The new feature column"Employment Type" is converted from the "Workclass" column. The new column contains values of 'private','government','self-employed' or 'without-pay'
4. Using Scipy, we performed correlation with the 'salary' column with the 'point biserial' method.That given the result that 'fnlwgt' and 'marital-status'

column are negatively correlated with the ‘salary’. So, we dropped those two columns.

5. Then splitted the whole dataset into a train-test split. As our dataset is Imbalanced, this is handled using SMOTE technique by oversampling the dataset.
6. After performing the above step the dataset is ready for training. In this step, Logistic Regression, Random Forest, Gradient Boosting, Decision Tree, SVM, XGBoost, Catboost, KNN are used.
7. After training all above models Catboost performed the best, and cross-validation is also done using the ‘RepeatedStratifiedKFold’ method. Which got the accuracy over 80%.
8. Then the model is saved using pickle file format for model deployment.
9. After the model was ready to deploy. For deployment Heroku platform is used.

4.1 Re-usability

The programming is done in such a way for this project that it should be reusable. So that anyone can add and contribute without facing any problems.

4.2 Application Compatibility

The difference module of this project is using Python as an interface between them. Each module has its own job to perform and it is the job of the Python to ensure the proper transfer of information.

4.3 Deployment

We have deployed this on cloud and also dockerized this.

5. Conclusion

The Adult Income Census prediction model will classify whether the person’s income will be >50K or <=50K..