High Level Document (HLD)

Adult Census Income Prediction

Version: 1.0

Date: 30/10/2021

Contents

Abstract

The prominent inequality of wealth and income is a huge concern around the world. The likelihood of diminishing poverty is one valid reason to reduce the world's surging level of economic inequality. Governments in different countries have been trying their best to address this problem and provide an optimal solution.

This study aims to show the usage of machine learning and data mining techniques in providing a solution to the income equality problem. The UCI Adult Dataset has been used for the purpose. Classification has been done to predict whether a person's yearly income in US falls in the income category of either greater than 50K Dollars or less than/equal to 50K Dollars category based on a certain set of attributes. The Gradient Boosting Classifier Model was deployed which clocked the highest accuracy of 86 %

The salary of a person varies with respect to various parameters, which can be education, country, domain etc. Salary is an important thing for the finance, retail or E commerce sector. In addition, It’s an important variable for the categorization of the customers with respect to their salary/financial situation.

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 things about this project.

2. General Description

2.1 Product Perspective

The Adult Census Income Prediction using Classifier based Machine Learning algorithms.

2.2 Problem statement

The problem of income inequality has been of great concern in the recent years. Making the poor better off does not seem to be the sole criteria to be in quest for eradicating this issue. People around the world believe that the advent of economic inequality is unacceptable and demands a fair share of wealth in the society. This model actually aims to conduct a comprehensive analysis to highlight the key factors that are necessary in improving an individual's income. Such an analysis helps to set focus on the important areas, which can significantly improve the income levels of individuals.

2.3 Proposed Solution

The solution here is a Classifier based Machine Learning model. It can be implemented by different regression Classifier (like Logistic, Random-Forest, K-NN, DecisionTree, SVC, NaveBayes, AdaBoost, GradientBoosting).Here First we are performing Data preprocessing step, in which feature engineering, feature selection, feature scaling steps are performed and then we are going to build model.

2.4 Technical Requirements

In this Project, the requirements to get Income Prediction through various platform. For that, in this project we are going to use different technologies. Here is 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 (AWS, Heroku).
- Cassandra database should be integrated in this project for any kind of user input.

2.5 Data Requirements

Data Requirement completely depend on our problem.

- For training and testing the model, we are using adult census dataset from Kaggle.
- From user we are taking following input:

    - **High school grade/College –** (Bachelors, HS-grad, 11th, Masters, 9th, Some-college, Assoc-acdm, Assoc-voc, 7th-8th, Doctorate, Prof-school, 5th-6th, 10th, 1st-4th, Preschool, 12th)
    - **Age -** numeric value (range between 15 to 95)
    - **Work Class** – (State-gov, Self-emp-not-inc, Private, Federal-gov, Local-gov, Self-emp-inc, Without-pay, Never-worked)
    - **Final Weight -** Numeric value (range between 13000 to 650000)
    - **Occupation –** (Adm-clerical, Exec-managerial, Handlers-cleaners, Prof-specialty, Other-service, Sales, Craft-repair, Transport-moving, Farming-fishing, Machine-op-inspct, Tech-support, Protective-serv, Armed-Forces, Priv-house-serv)
    - **Capital Loss** - Numeric Value (must be in range between 0 to 4500)
    - **Relationship** – (Not-in-family, Husband, Wife, Own-child, Unmarried, Other-relative)
    - **Capital Gain** – Numeric Value (range between 0 to 100000)

- **Marital Status** – (Never-married, Married-civ-spouse, Divorced, Married-spouse-absent, Separated, Married-AF-spouse, Widowed)
- **Sex** – Male/Female
- **Race** – (White, Black, Asian-Pac-Islander, Amer-Indian-Eskimo, Other)
- **Hours Per Week** - Numeric range (1-100)
- **Country** - (north_America, central_America, central_Asia, South, west_Europe, East_Asia, east_Europe, south_America, China)

2.6 Tools Used

- PyCharm is used as IDE.

- For visualization of the plots, Matplotlib, Seaborn are used.
- Azure for deployment of the model.
- Cassandra to retrieve, insert, delete, and update the database.
- Front End development is done using HTML/CSS, Flask is used for backend development and for API development.
- GitHub as version control system.

2.7 Constraints

The Adult Census Income Prediction system must be user friendly, errors free and user should not be required to know any of the back-end working

3.0 Design Details

DETAILS 3.1

Process Flow

Data Collection -> Data Cleaning -> Feature Engineering

Handling Categorical -> Variable Feature Selection -> Train-Test Split

-> Feature Scaling Model Training -> Hyper Parameter

-> Tuning Model Testing -> Model Testing

-> ***Model Deployment***

3.2 Event Log

Make Prediction -> Display Predicted Result

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, class.
- By logging, we can monitor every insertion, every flow of data in database.
- By logging we are monitor every step, which may create problem, or every step, which is important in file system.
- We have designed logging in such a way that system should not hang even after so many logging's, so that we can easily debug issues which may arises during process flow.

3.3 Error Handling

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

4. Performance

Solution of Adult Census Income Prediction is used to predict the salary of the person; it should be as accurate as possible so that it should give accurate category prediction. That is why before building this model we followed complete process of Machine Learning. Here are summary of complete process:

1. First we cleaned our dataset properly by removing all null value, invalid and duplicate values present in dataset.
2. Then we categorized the salary (>50 as ‘1’ and <=50 as ‘0’) and performed feature extraction, in which I extracted country and reduced it to lesser categorical values
3. Then we performed mapping for parameters, assigning numeric values accordingly to various categories

4. Then I handled categorical variable by performing One-Hot encoding.
5. Then I split the whole data set train-test split. After that, I performed scaling on X_train and X_test.
6. . After performing above step I was ready for model training. In this step, I trained my dataset on different Classifier Learning algorithm (Logistic, Random-Forest, K-NN, DecisionTree, SVC, NaveBayes, AdaBoost, GradientBoosting). After training the dataset on different algorithms, I got highest accuracy of 80% on RadomForestClassifier.
7. Then I applied hyper-parameter tuning on all models which I have described above. Here also I got highest accuracy of 86% on test dataset by same GradientBoostingClassifier
8. After that I saved my model in pickle file format for model deployment.
9. After that my model was ready to deploy. I deployed this model on various cloud storage (AWS and heroku) and dockerize this model.

4.1 Re-usability

We have done programming of this project in such a way that it should be reusable. Therefore, anyone can add and contribute without facing any problems.

4.2 Application Compatibility

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

4.3 Resource Utilization

In this project, when any task is performed, it will likely that the task will use all the processing power available in that particular system until it's job finished. By keeping this in mind, In this project we have used the concept of multi-threading.

4.4 Deployment

We have deployed this on cloud and dockerized this.

4.5 User Interface

We have created an UI for user by using HTML and CSS.

5. Conclusion

The Adult Census Income Prediction model will be predicting the income of the population and analyzing the factors, which strongly affect the income. Giving suggestions based on the result obtained which level of qualification can lead to a higher income and people of which age group are earning more.

In addition to it, suggestions could be given based on the predictions to students who are in need to pursue higher education or people who are spending less time in the workplace.