

This study aims to show the usage of machine learning and data mining techniques in providing a solution to the income equality problem

High-Level Design

Adult Census Income Prediction

Saakar Sengar And Shreyansh Jain

Contents

Document Version Control Abstract

1. Introduction
 - 1.1 Why this High Level Design Document?
2. General Description
 - 2.1 Product Perspective
 - 2.2 Problem Statement
 - 2.3 Proposed Solution
 - 2.4 Technical Requirements
 - 2.5 Data Requirements
 - 2.6 Tools Used
 - 2.7 Constraints
3. Design Details
 - 3.1 Process Flow
 - 3.2 Deployment Process
 - 3.2 Event Log
 - 3.3 Error Handling
4. Performance
 - 4.1 Re-usability
 - 4.2 Application Compatibility
 - 4.3 Resource Utilization
 - 4.4 Deployment
 - 4.5 User Interface
5. Conclusion

Abstract

The prominent inequality of wealth and income is a huge concern especially in the United States. The likelihood of diminishing poverty is one valid reason to reduce the world's surging level of economic inequality. The principle of universal moral equality ensures sustainable development and improve the economic stability of a nation. 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 equal to 50K Dollars category based on a certain set of attributes

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, classified the income category of either greater than 50K Dollars or less equal to 50K Dollars category of the person by using classification based Supervised 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 Supervised Machine Learning model. It can be implemented by different classification algorithms (like Logistic Regression, Random Forest Classification, Decision Tree Classification, SVC, Xg-boost Classifier and so on.). Here first, we are performing Data pre-processing 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 (Azure, AWS, GCP, 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 Income Prediction dataset from Kaggle.
- ❖ From user we are taking following input :
 - **Age** : an integer value – user age
 - **Capital Gain** : an integer value b/w [0-99999]
 - **Capital Lose** : an integer value b/w [0-3456]
 - **Hours per Week** : an integer value b/w [1-99]
 - **Education** : Preschool, 1st - 4th, 5th - 6th, 7th - 8th, 9th, 10th, 11th, 12^{ht}, HS-grad, Somecollege, Bachelors, Masters, Assoc-voc, Assoc-acdm, Prof-school, Doctorate.
 - **Work-Class** : Private, Self-emp-not-Inc., Local-gov, State-gov, Self-emp-Inc., Federalgov, Without-pay, Never-worked
 - **Marital Status** : Married-AF-spouse, Married-civ-spouse, Married-spouse-absent, Never-married, Separated, Widowed
 - **Occupation** : Adm-clerical, Armed-Forces, Craft-repair, Exec-managerial, Farmingfishing, Handlers-cleaners, Machine-op-inspct, Other-service, Priv-house-serv, Profspecialty, Protective-serv, Sales, Tech-support, Transport-moving.
 - **Race** : Asian-Pac-Islander, Black, Other, White
 - **Gender** : Male, Female
 - **Relationship** : Not-in-famil ,Other-relative, Own-child, Unmarried, Wife
 - **Native Country** : ' United-States', ' Cuba', ' Jamaica', ' India', ' Mexico', ' Puerto-Rico', ' Honduras', ' England', ' Canada', Germany', ' Iran', ' Philippines', ' Italy', ' Poland', ' Columbia', ' Cambodia', ' Thailand', ' Ecuador', ' Laos', ' Taiwan', ' Haiti', ' Portugal', ' Dominican-Republic', ' El-Salvador', ' France', ' Guatemala', ' China', ' Japan', ' Yugoslavia', ' Peru', ' Outlying-US(Guam-USVI-etc)', Scotland', ' Trinidad&Tobago', ' Greece', ' Nicaragua', ' Vietnam', ' Hong', Ireland', ' Hungary', ' Holand-Netherlands'

2.6 Tools Used



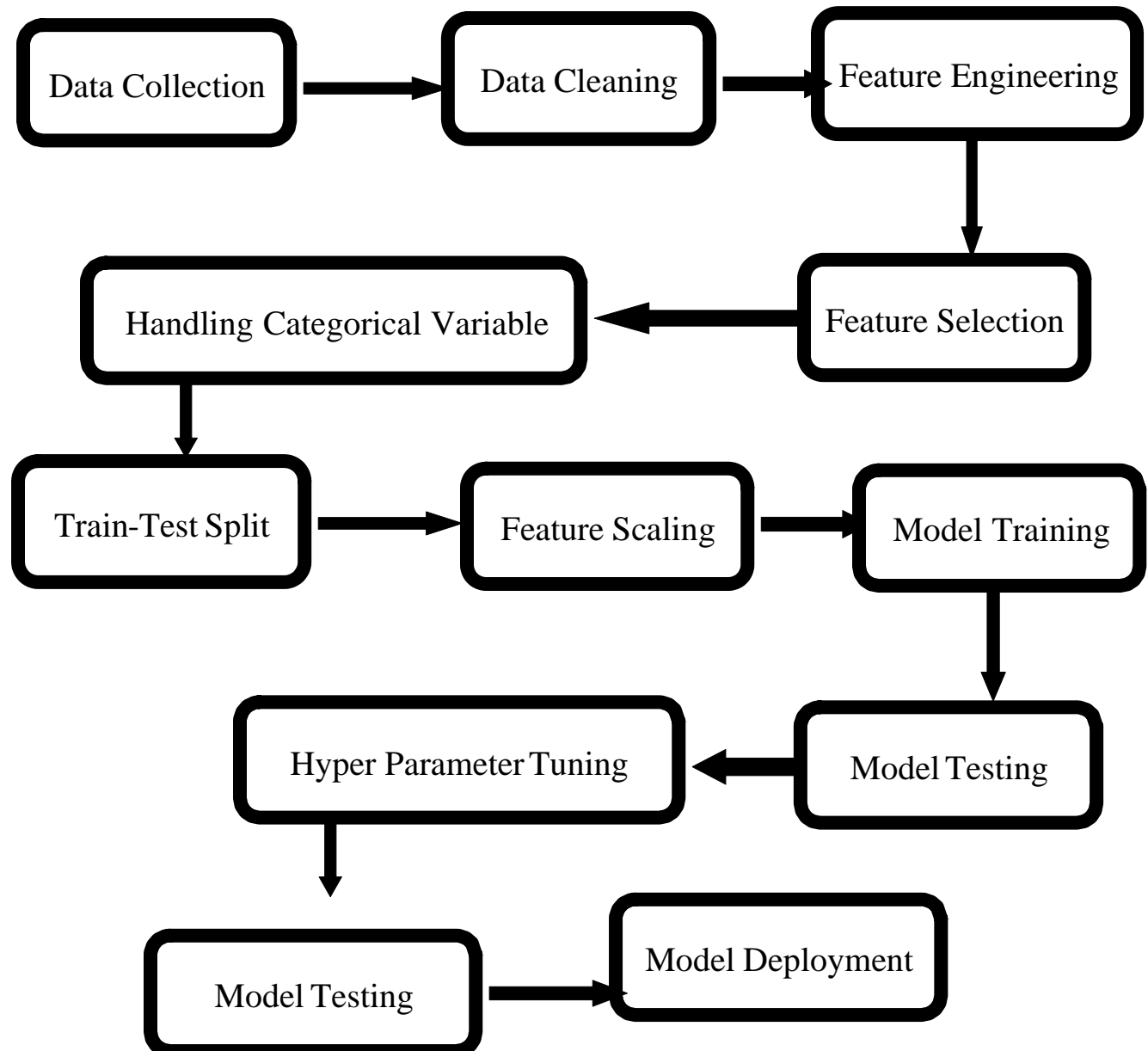
Tools	Uses
Python	high-level computer programming language used to develop the project
Py-Charm	an integrated development IDE used in computer programming, for the Python language
Pandas	Open source data analysis and manipulation tool, for the Python programming language.
Numpy	Python library used for working with arrays
Matplotlib/Seaborn	For data visualization and graphical plotting library for Python
Scikit-Learn	Machine learning library used for the Python programming language. It features various classification, regression and clustering algorithms
Flask	a web framework, it's a Python module that lets you develop web applications easily
HTML/CSS	Are two of the core technologies used for building Web pages. HTML provides the structure of the page, CSS the layout for a variety of devices.
Heroku	Is used as a platform as a service (PaaS) that build, run, and operate applications entirely in the cloud
GitHub	web-based interface that used Git for the open source version control system

2.7 Constraints

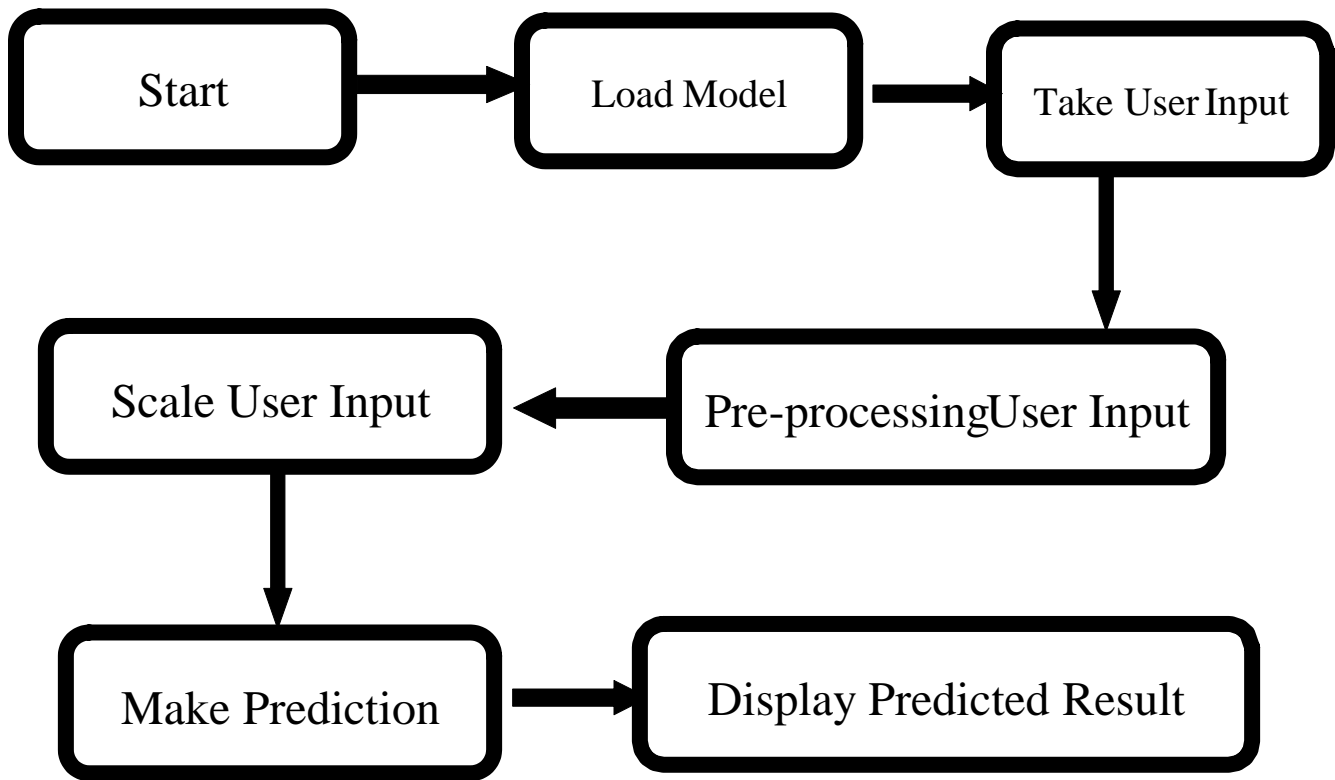
The Adult Census Income Prediction system must be user friendly, errors free and users should not be required to know any of the back-end working.

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, 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.

3.4 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

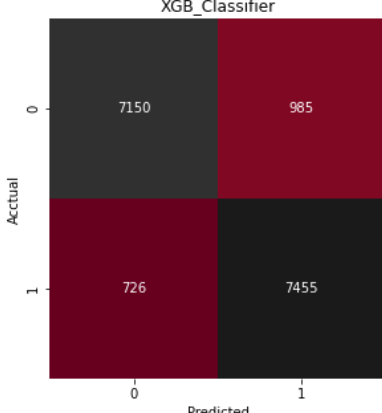
Solution of Adult Census Income Prediction is used to classified into the $>50K$ group or $\leq 50K$ group in advance, so it should be as accurate as possible so that it should give as much as possible accurate Income prediction.

That's 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 extra space present and duplicate value present in dataset.
2. Separate the dependent and independent variables.
3. according to first null values analysis there is no any missing value present in dataset, but take insight look in data we can see that a symbol '?' present in dataset which means that this is present at the place of missing value, so we need to replace all these '?' with 'null' values
4. there is missing values are present in columns 'work-class', 'occupation' & 'country', all these columns are categorical columns so we impute the missing values with Mode.
5. Then we compute all the outliers present in dataset and handle them.
6. Then we handled categorical variable by performing One-Hot encoding rather than country columns, in country 98% of data refers to United States, so we differentiate data into two entities USA and Non-USA.
7. The data is highly imbalanced then we balanced the dataset by using standard scaler.
8. Then we split the whole data set train-test split. And split into X_{train} , X_{test} , y_{train} and y_{test} .

9. After performing above step I was ready for model training. In this step, I trained my dataset on different classification based supervised Machine Learning Algorithm (Logistic Regressions, Random-Forest Classification, XGBoost Classifier and Gaussian NB). After training the dataset on different algorithms I got highest accuracy of 86% on XGBoost Classifier
10. After that I applied hyper-parameter tuning on all model which I have described above. Here also I got highest accuracy of 90% on test dataset by same XGBoost Classifier.

		precision	recall	f1-score	support	
	0	0.91	0.88	0.89	8135	
	1	0.88	0.91	0.90	8181	
	accuracy			0.90	16316	macro
avg	0.90	0.90	0.90	16316	weighted avg	
0.90	0.90	0.90	16316			



Confusion matrix for XGB_Classifier:

	Predicted 0	Predicted 1
Actual 0	7150	985
Actual 1	726	7455

11. After that I saved my model in pickle file format for model deployment.
12. After that my model was ready to deploy. I deployed this model on Heroku which is used as a platform as a service (PaaS) to build, run, and operate applications entirely in the cloud.

4.1 Re-usability

We have done programming of this project in Modular Fashion in which various classes are made so that it should be reusable. So that 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 module 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 on Heroku platform as a service.



4.5 User Interface

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

5. Conclusion

The Adult Income Prediction model will predict the income category of a user based on various attributes. The model will be trained on all the different attributes of users which help the model to give accurate predictions if the salary of the individual is more than 50K or not.

Thank You