Architecture Design

Credit Card Default Prediction

Document Control

Version	Date	Author	Comments
1	18.08.2024	Abhishek Upadhyay	

Index

Abstract

Machine Learning is a category of algorithms that allows software applications to become more accurate in predicting outcomes without being explicitly programmed. The basic premise of machine learning is to build models and employ algorithms that can receive input data and use statistical analysis to predict an output while updating outputs as new data becomes available. These models can be applied in different areas and trained to match the expectations of management so that accurate steps can be taken to achieve the organization's target. In this project, we will estimate the amount of insurance premium on the basis of personal health information. Taking various aspects of a dataset collected from people, and the methodology followed for building a predictive model.

1. Introduction

1.1 What is Architecture Design?

The goal of Architecture Design (AD) is to give the internal design of the actual program code for the `Insurance Premium Prediction`. AD describes the class diagrams with the methods and relation between classes and program specification. It describes the modules so that the programmer can directly code the program from the document.

1.2 Scope

Architecture Design (AD) is a component-level design process that follows a step-by-step refinement process. This process can be used for designing data structures, required software, architecture, source code, and ultimately, performance algorithms. Overall, the data organization may be defined during requirement analysis and then refined during data design work. And the complete workflow.

1.3 Constraints

We only predict the expected estimating cost of expenses customers based on some personal health information.

2. Technical Specification

2.1 Dataset

The dataset containing verified historical data, consisting of the aforementioned information

Architecture Design

There are 1001 rows and 24 columns in the given datasets . By Using these data we are build a model that should Predict that the default payment next month this is a binary class classification problem.

	SEX	EDUCATION	MARRIAGE	AGE	PAY_0	PAY_2	PAY_3	PAY_4	PAY_5	...	BILL_AMT4	BILL_AMT5	BILL_AMT6	PAY_AMT1	PAY_AMT2	PAY_AMT3	PAY_AMT4	PAY_AMT5	PAY_AMT6	defau payme ne mon
	1	2	1	57	-1	0	-1	0	0	...	20940	19146	19131	2000	36681	10000	9000	689	679	
0	1	1	2	37	0	0	0	0	0	...	19394	19619	20024	2500	1815	657	1000	1000	800	
1	1	1	2	29	0	0	0	0	0	...	542653	483003	473944	55000	40000	38000	20239	13750	13770	
2	2	2	2	23	0	-1	-1	0	0	...	221	-159	567	380	601	0	581	1687	1542	
3	2	3	1	28	0	0	2	0	0	...	12211	11793	3719	3329	0	432	1000	1000	1000	
4	nc																			

The data set consists of various data types from integer to floating to object as shown in Fig

```
1  data.info()
```

Python

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 30000 entries, 0 to 29999
Data columns (total 24 columns):
LIMIT_BAL                   30000 non-null int64
SEX                         30000 non-null int64
EDUCATION                   30000 non-null int64
MARRIAGE                    30000 non-null int64
AGE                         30000 non-null int64
PAY_0                       30000 non-null int64
PAY_2                       30000 non-null int64
PAY_3                       30000 non-null int64
PAY_4                       30000 non-null int64
PAY_5                       30000 non-null int64
PAY_6                       30000 non-null int64
BILL_AMT1                   30000 non-null int64
BILL_AMT2                   30000 non-null int64
BILL_AMT3                   30000 non-null int64
BILL_AMT4                   30000 non-null int64
BILL_AMT5                   30000 non-null int64
BILL_AMT6                   30000 non-null int64
PAY_AMT1                    30000 non-null int64
PAY_AMT2                    30000 non-null int64
PAY_AMT3                    30000 non-null int64
PAY_AMT4                    30000 non-null int64
PAY_AMT5                    30000 non-null int64
PAY_AMT6                    30000 non-null int64
default payment next month  30000 non-null int64
dtypes: int64(24)
memory usage: 5.5 MB
```

Type your text

Various factors important by statistical means like mean, standard deviation, median, count of values and maximum value, etc. are shown below for numerical attributes

Preprocessing of this dataset includes doing analysis on the independent variables like checking for null values in each column and then replacing or filling them with supported appropriate data types so that analysis and model fitting is not hindered from their way to accuracy. Shown above are some of the representations obtained by using Pandas tools which tell about variable count for numerical columns and model values for categorical columns. Maximum and minimum values in numerical columns, along with their percentile values for median, play an important factor in deciding which value to be chosen at priority for further exploration tasks and analysis. Data types of different columns are used further in label processing and a one-hot encoding scheme during the model building.

2.2 Logging

We should be able to log every activity done by the user

- The system identifies at which step logging require.
- The system should be able to log each and every system flow.
- The system should be not be hung even after using so much logging. Logging just because we can easily debug issuing so logging is mandatory to do.

2.3 Deployment

For the hosting of the project, we will use Render.

3. Technology Stack

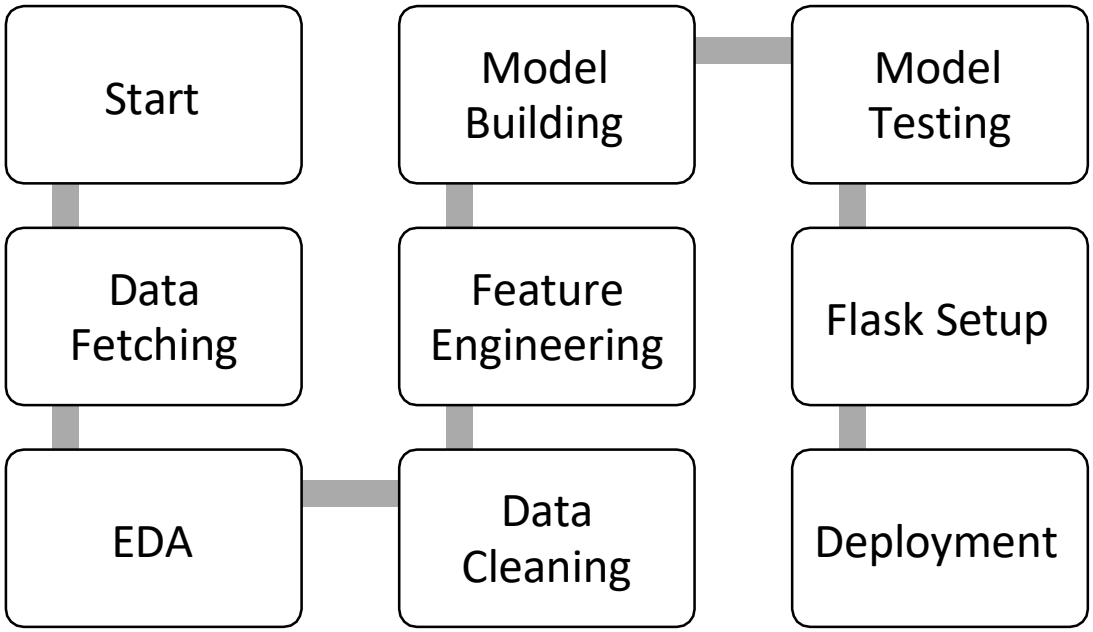
Front End	HTML/CSS
Backend	Python/ Flask
Deployment	Render

4. Proposed Solution

We will use performed EDA to find the important relation between different attributes and will use a machine-learning algorithm to estimate the cost of expenses. The client will be filled the required feature as input and will get results through the web application. The system will get features and it will be passed into the backend where the features will be validated and preprocessed and then it will be passed to a hyperparameter tuned machine learning model to predict the final outcome.

5. Architecture

Type your text



5.1 Data Gathering

Data source: https://www.kaggle.com/datasets/overload10/adult-census-dataset

Dataset is stored in .csv format.

Type your text

5.2 Raw Data Validation

After data is loaded, various types of validation are required before we proceed further with any operation. Validations like checking for zero standard deviation for all the columns, checking for complete missing values in any columns, etc. These are required because the attributes which contain these are of no use. It will not play role in contributing to the estimating cost of the premium.

5.3 Exploratory Data Analysis

Visualized the relationship between the dependent and independent features. Also checked relationship between independent features to get more insights about the data.

5.4 Feature Engineering

After pre-processing standard scalar is performed to scale down all the numeric features. Even one hot encoding is also performed to convert the categorical features into numerical features. For this process, pipeline is created to scale numerical features and encoding the categorical features.

5.5 Model Building

After doing all kinds of pre-processing operations mention above and performing scaling and encoding, the data set is passed through a pipeline to all the models, Logistic Regression, Decision tree, Random Forest, Gradient boost, KNN and XGBoost classification using EvalML. It was found that Gradient boosting classifier performs best with the 84.87050875217524 accuracy value.

Type your text

5.6 Model Saving

Model is saved using pickle library in pickle` format.

Type your text

5.7 Flask Setup for Web Application

After saving the model, the API building process started using Flask. Web application creation was created in Flask for testing purpose. Whatever user will enter the data and then that data will be extracted by the model to estimate the premium of insurance, this is performed in this stage.

Type your text

5.8 GitHub

The whole project directory will be pushed into the GitHub repository.

5.9 Deployment

The project was deployed from GitHub into the Heroku platform.

6. User Input / Output Workflow.