

Credit Card Default Prediction

A series of several thin, white, parallel diagonal lines extending from the bottom right towards the top right of the slide, adding a modern, geometric design element.

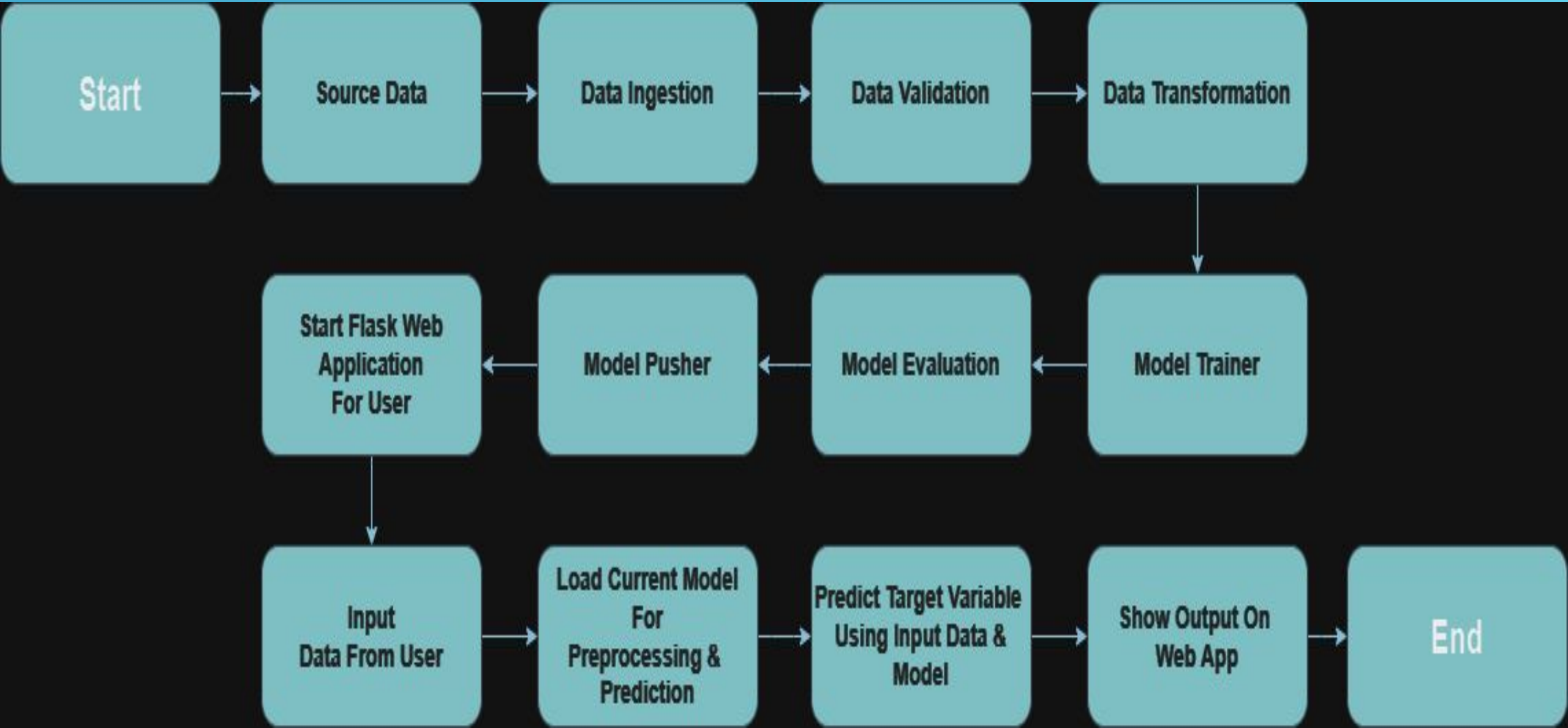
Objective:

Develop a predictive model for credit card default prediction to identify customers at risk of payment default and enhance risk assessment for financial institutions.

Benefits:

- Early identification of credit card defaults, facilitating proactive risk management and mitigation.
- Enhanced understanding of the customer base, leading to more personalized and targeted credit services.
- Streamlined resource management, optimizing operational efficiency in credit risk assessment.
- Manual inspection available for cases of potential default, ensuring thorough investigation and appropriate action.

Architecture



Data Validation and Data Transformation :

- ❑ Number of Columns – Validation of number of columns present in the files, and if it doesn't match, then file is not accepted for further process.
- ❑ Name of Columns - The name of the columns is validated and should be the same as given in the schema file. If not, then file is not accepted for further process.
- ❑ Data type of columns - The data type of columns is given in the schema config file. It is validated when we insert the files into further pipeline process. If the datatype is wrong, then file is not accepted for further process..
- ❑ Null values in columns - If any of the columns in a file have all the values as NULL or missing, values are preprocessed using preprocessing object.

Model Training:

□ Data Import From Client :

The accumulated data from client is exported as csv , compressed in zip format

□ Data Preprocessing

- Performing EDA to get insight of data like identifying distribution , outliers.
- Check for null values in the columns. If present impute the null values.
- Perform Standard Scalar to scale down the values.

□ Classification –

- In the classification phase, various machine learning algorithms are employed to train a model on the preprocessed data. The objective is to accurately classify instances into predefined categories or classes based on their features. To determine the best-performing algorithm and ensure optimal model performance, evaluation metrics such as accuracy and F1 score are utilized.
- The RandomForestClassifier model is trained over preprocessed data, and the model is saved for further use in prediction

□ Model Selection –

After model training , we find the best model trained from parameters in model.yaml file. Which ever algorithms is performing best , we will choose that algorithm for prediction purpose.

Prediction:

- We have create web application to take input from user, for prediction
- We perform data pre-processing techniques on it.
- We use current model in production for prediction purpose
- Once valid input is given, after clicking prediction button , prediction will be displayed on the web page.

Q & A:

Q1) What's the source of data?

The data for training is provided by the client in csv format compressed in zip format

Q 2) What was the type of data?

The data consist of all integer numerical values.

Q 3) What's the complete flow you followed in this Project?

Refer slide 4th for better Understanding

Q 4) After the File validation what you do with incompatible file or files which didn't pass the validation?

We don't use these files for training purpose, we need correct data for model training.

Q 5) How logs are managed?

We are using different logs as per the steps that we follow in validation and modeling like data ingestion, data validation ,data transformation , model trainer , model evaluation , model pusher , pipeline, etc.

Q 6) What techniques were you using for data pre-processing?

- ▶ Removing unwanted attributes
- ▶ Checking data drift
- ▶ Checking and changing Distribution of continuous values
- ▶ Cleaning data and imputing if null values are present.
- ▶ Scaling the data

Q 7) How training was done or what models were used?

- ▶ Data is divided into training and testing set , using stratify method.
- ▶ The scaling was performed over training and validation data.
- ▶ Algorithms like LogisticRegression, Decision Tree , RandomForestClassifier , SVC were used based on the F1 score final model was used for prediction and we saved that model .

Q 8) How Prediction was done?

Web application created for user , to input data. Data collected will go through , preprocessed object.

After data is preprocessed, trained model will generate prediction for that particular record.

Prediction will be shown on the same web app.

