DPR

Insurance Premium Prediction

Revision Number – 1.0

Last Date of Revision – 18/05/2024

Document Version Control

Date	Version	Description	Author
16-05-2024	1.0	Introduction	Sakshi Choube
17-05-2024	1.1	Deployment	Sakshi Choube
18-05-2024	1.2	Q and A	Sakshi Choube

Contents

INTRODUCTION

Why this DPR Documentation?

The main purpose of this DPR documentation is to add the necessary details of the project and provide the description of the machine learning model and the written code. This also provides the detailed description on how the entire project has been designed end-to-end.

Key points:

Describes the design flow
Implementations Software
requirements Architecture of the
project
Non-functional attributes like:Reusability
Portability Resource
utilization

1 Description

1.1 Problem Perspective

1. Data Quality and Generalization: Data limitations and potential overfitting could affect the accuracy and generalizability of predictive models.
2. Privacy and Regulations: Ethical concerns and evolving insurance regulations may pose challenges in handling personal data and ensuring compliance.

- Model Transparency and Adoption: Complex models might lack transparency, hindering user understanding and acceptance.
- External Factors: Economic and environmental influences on insurance markets may not be fully considered in predictions.

1.2 Problem Statement

The goal of this project is to give people an estimate of how much they need based on their individual health situation. After that, customers can work with any health insurance carrier and its plans and perks while keeping the projected cost from our study in mind. This can assist a person in concentratingon the health side of an insurance policy rather than the ineffective part.

1.3 Proposed Solution

Our solution involves developing a personalized premium prediction model that utilizes advanced machine learning techniques. It will consider individual factors like age, smoker, sex to generate precise premium estimates. We will ensure data quality and privacy compliance, create a user-friendly interface, and continuously update the model to adapt to changing regulations and market conditions. Additionally, educational efforts will promote awarenessand adoption among consumers and industry stakeholders.

1.4 Solution Improvements

Improvements the solution for leveraging previous insurance data, fostering cross-functional collaboration, integrating external data sources, continuous monitoring, machine learning models, and customer feedback. By implementing these enhancements, the solution becomes more comprehensive, adaptable, and capable of addressing evolving challenges, enabling individuals to stay informative and responsive about the premiums.

1.5 Technical Requirements

There are not any hardware needs needed for victimization this application, the user should have an interactive device that has access to the web and should have the fundamental understanding of providing the input. And for the backend half the server should run all the package that's needed for the process and provided information to show the results.

1.6 Tools Used

- □ Python 3.8 is employed because the programming languageand frame
  works like Numpy, Pandas, Sklearn, Flask, Streamlit and alternative modules for building the model.
- □ Visual Studio Code is employed as IDE.
- □ Front end development is completed victimization HTML/CSS
- □ Flask is employed for each information and backend readying
- □ GitHub is employed for version management
- □ Streamlit Cloud and localhost is used for Deployment


scikit
learn
GitHub
Pandas
Streamlit
NumPy
Flask

1.7 Data Requirements

The information needed for the project is completely supported by the matter statement. The dataset is available on Kaggle in the form of an Excel spreadsheet (.xlsx). Since the main theme of the project is to gain experience with real-time issues, we will import the data into the Oracle database and then export it into CSV format.

1.8 Data Gathering from Main Source

The data for ~~the current project is bei~~ng gathered from the Kaggle dataset. This dataset serves as the primary source for our project's data analysis and premium prediction tasks.

1.9 Data Description

We have train (1070) and test (268) data set, train data set has both input and output

Columns Are:

Age: Age of the insured individuals, a key factor in premium calculation due to its influence on health risk.
Sex: Gender of the insured individuals, impacting premiums based on gender-specific health risks. BMI (Body Mass Index): Measure of body weight relative to height, influencing premiums based on health implications.
Children: Number of dependents covered, affecting policy costs due to family size.
Smoker: Smoking status (yes/no), a significant factor in premium pricing due to health risks associated with smoking.
Region: Geographic location of the insured, which can affect healthcare costs and insurance pricing.
Expenses: Actual medical expenses incurred, providing insight into healthcare costs and utilization for premium calculation.

age	sex	bmi	children	smoker	region	expenses
19	female	27.9	0	yes	southwest	16884.92
18	male	33.8	1	no	southeast	1725.55
28	male	33	3	no	southeast	4449.46
33	male	22.7	0	no	northwest	21984.47
32	male	28.9	0	no	northwest	3866.86
31	female	25.7	0	no	southeast	3756.62
46	female	33.4	1	no	southeast	8240.59
37	female	27.7	3	no	northwest	7281.51
37	male	29.8	2	no	northeast	6406.41
60	female	25.8	0	no	northwest	28923.14
25	male	26.2	0	no	northeast	2721.32
62	female	26.3	0	yes	southeast	27808.73
23	male	34.4	0	no	southwest	1826.84
56	female	39.8	0	no	southeast	11090.72
27	male	42.1	0	yes	southeast	39611.76
19	male	24.6	1	no	southwest	1837.24
52	female	30.8	1	no	northeast	10797.34
23	male	23.8	0	no	northeast	2395.17
56	male	40.3	0	no	southwest	10602.39
30	male	35.3	0	yes	southwest	36837.47
60	female	36	0	no	northeast	13228.85
30	female	32.4	1	no	southwest	4149.74

1.10 Data Ingestion

The cornerstone of our data-driven project was laid through a systematic process of data acquisition and ingestion. Leveraging Kaggle, a renowned platform esteemed for its high-quality datasets, we pinpointed and obtained the essential data required for our insurance price prediction endeavor. This dataset, pivotal to our objective of accurate price forecasting, was meticulously downloaded and securely stored within our local system infrastructure.

Subsequently, we initiated the data ingestion phase, seamlessly integrating the dataset into our project's data pipeline. This meticulous approach ensures that our project is built upon a solid foundation, laying the groundwork for robust and precise insurance price prediction models and analysis.

1.11 Data Transformation

Steps performed in pre-processing are:

- First read data from Artifact folder
- Checking unnecessary columns
- One column has product id which is unique for every product so I deleted that column.
- Checked for null values
- there are too many null values are present in two columns that's why I deleted them
- Performed one-hot encoder on categorical columns.
- Scaling is performed for needed information.
- And, the info is prepared for passing to the machine learning formula

1.12 Modelling

After pre-processing the information, we envision the dataset, extracting all the necessary insights. Although the insights are initially scattered, we proceed with modeling using various machine learning algorithms to ensure comprehensive coverage of all possibilities. Eventually, Gradient Boosting emerged as the top performer, affirming its effectiveness in our predictive modeling endeavors.

1.13 UI Integration

Both CSS and HTML files are being created and are being integrated with the created machine learning model. All the required files are then integrated to the app.py file and tested locally.

1.14 Data from User

The data from the user is retrieved from the created HTML web page.

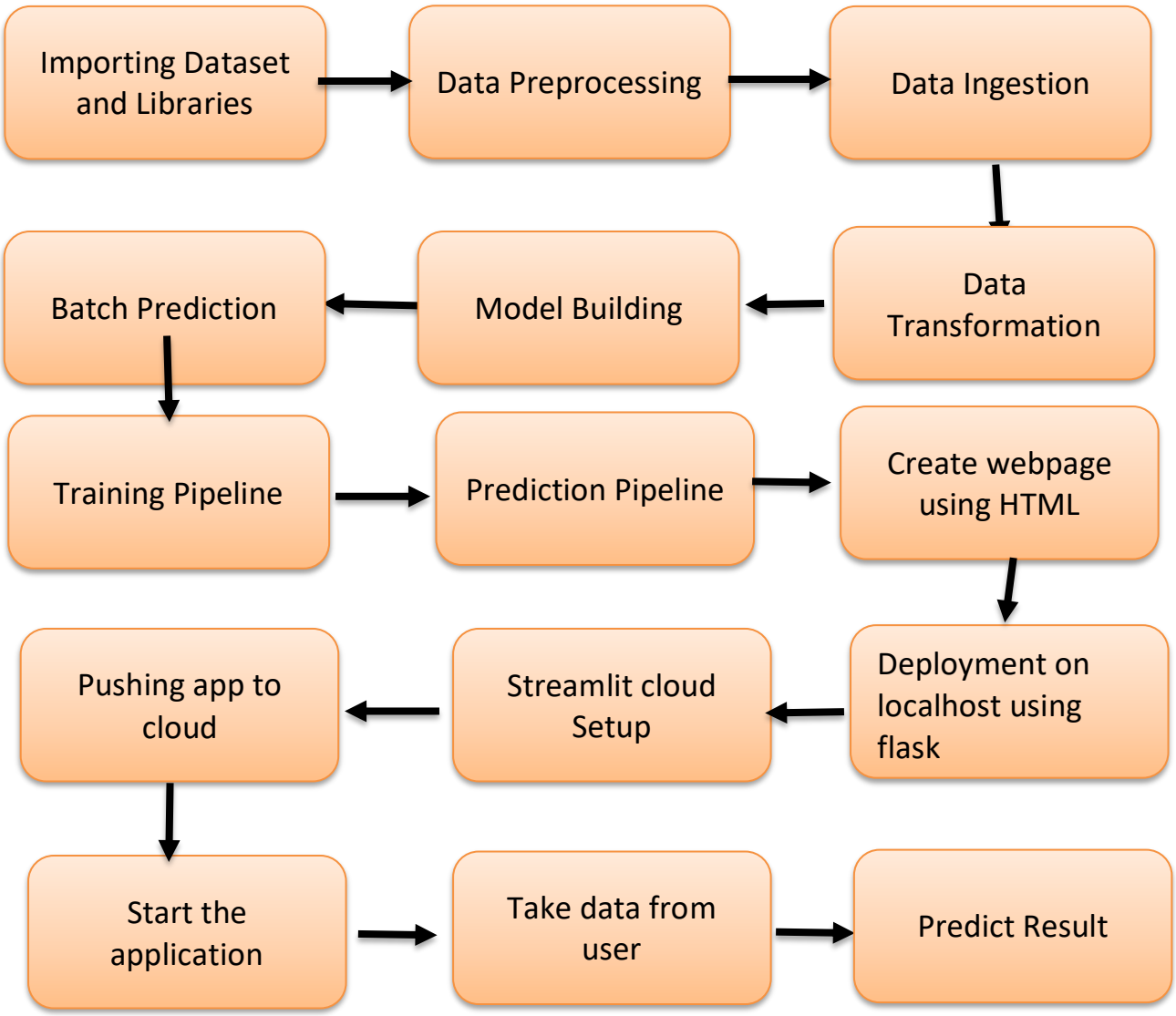
1.15 Rendering the Results

The data sent for the prediction is then rendered to the web page.

1.16 Deployment

The tested model is then deployed to Streamlit Cloud. So, users can access the project from any internet devices.

Modelling process & Deployment



Conclusion

Insurance Premium Prediction represents a crucial step towards informed insurance purchases. By harnessing data analysis and machine learning, it brings transparency and fairness to premium calculation, empowering usersto make confident, personalized choices. As this journey continues, it promises to reshape the insurance landscape, ensuring individuals and businesses can purchase premiums with clarity and confidence.

Q & A:

Q1) What's the source of data?
Ans-The data for training is provided by the client in multiple batches and each batch contain multiple files.

Q 2) What was the type of data?
Ans-The data was the combination of numerical and Categorical values.

Q 3) What's the complete flow you followed in this Project?
Ans-Refer Page no 11 for better Understanding.

Q 4) After the File validation what you do with incompatible file or files which didn't pass the validation?
Ans-Files like these are moved to the Achieve Folder and a list of these files has been shared with the client and we removed the bad data folder.

Q 5) How logs are managed?
Ans- We are using different logs as per the steps that we follow in validation and modeling like File validation log, Data Insertion, Model Training log, prediction log etc.

Q 6) What techniques were you using for data pre-processing?
Ans-Removing unwanted attributes Cleaning data and imputing if null values are present. Converting categorical data into numeric values.

Q 7) How training was done or what models were used?
Ans-Before dividing the data in training and validation set, we performed pre-processing over the data set and made the final dataset. As per the dataset training and validation data were divided. Algorithms like SVR ,Decision Tree, Random Forest, Gradient Boosting were used based on therecall, final model was used on the dataset and we saved that model.

Q 8) How Prediction was done?

Ans-The testing files are shared by the client. We Performed the same life cycle on the provided dataset. Then, on the basis of dataset, model is loaded and prediction is performed. In the end we get the accumulated data ofpredictions.

Q 9) What are the different stages of deployment?

Ans-First, the scripts are stored on GitHub as a storage interface.
The model is first tested in the local environment.
After successful testing, it is deployed on Streamlit Cloud.

DPR