2433 project

Qian Zhang
NYU ID: N19756113 Net ID: QZ2570

Xinge Zhang
NYU ID: N1083729 Net ID: XZ4513

December 2023

1 Introduction

The project is about creating a database and interactive system for an insurance company. This initiative aims to revolutionize the way the company manages its data, streamlines its processes, and interacts with customers. In the modern insurance sector, where data accuracy and efficient customer service are paramount, the need for a robust and user-friendly database system cannot be overstated. Our project focuses on developing a comprehensive solution that addresses these needs while ensuring scale ability, security, and compliance with industry standards.

The objective of this project is to develop a database application that not only stores and manages vast amounts of insurance-related data but also provides an interactive interface for both employees and clients. This interface is designed to facilitate easier access to policy information, claim processing, and customer support, thereby enhancing overall operational efficiency and customer satisfaction.

In the subsequent sections of this report, we will delve into the specific requirements that drove the development of this project, the methodologies employed in the system's design and implementation, and the various technologies and tools utilized. Additionally, we will outline the testing strategies adopted to ensure the reliability and effectiveness of the application, discuss the security measures implemented, and detail the deployment process and maintenance plans.

By the end of this report, it will be evident how this project not only meets the current demands of the insurance company but also positions it to adapt to future challenges and opportunities in the dynamic landscape of the insurance industry.

2 Business Case

2.1 Brief overview of the need of the application

2.1.1 Data Management and Integrity

The insurance sector generates vast amounts of data, including customer records, policy details, and claims information. Current systems struggle to manage this effectively. Inaccurate data can lead to significant problems, including incorrect policy quotations, claim processing errors, and regulatory non-compliance. Quick and reliable access to data is essential for timely decision-making and customer service The ability to integrate data from various sources and departments is crucial for a holistic view of operations.

2.1.2 Customer Service Enhancement

Current systems may not provide the speed necessary for prompt customer responses, leading to dissatisfaction. There's a growing demand for self-service options where customers can access their policy details, file claims, and get

instant responses. Customers expect personalized services based on their history and preferences, which current systems might not support adequately.

2.1.3 Compliance and Security

The insurance industry is heavily regulated, and non-compliance can lead to severe penalties. A new system can ensure adherence to these regulations more effectively. With increasing cyber threats, a robust system with enhanced security features is crucial to protect sensitive customer data. The ability to track and audit changes in data is vital for regulatory purposes and for internal checks.

2.1.4 Market Competitiveness and Growth

To remain competitive, the company needs to leverage technology effectively, especially in an industry that is rapidly adopting digital solutions. As the company grows, the system should be able to handle increased data and transaction volumes without performance degradation. A modern system can open up new possibilities, such as data analysis for better risk assessment and tailored insurance products.

2.1.5 Technological Advancements and Innovations

Current systems may be based on outdated technology, limiting their ability to adapt to new trends like AI and big data analysis. With the increasing use of mobile devices, the lack of mobile-friendly access to services is a significant drawback. The ability to integrate with IoT devices, wearable tech, and telematics for more advanced insurance models.

2.2 Cost-Benefit Analysis for the Database and Interactive System

2.2.1 Development Costs

Software Costs: Expenses related to the purchase or licensing of DBMS software, development tools, security software, and other necessary software components.
Hardware Costs: If not using cloud infrastructure, costs for servers, storage devices, and networking equipment.
Cloud Service Fees: Ongoing costs for cloud hosting services, if applicable.
Labor Costs: Expenses for the development team, including analysts, developers, project managers, and QA testers.
Consultancy Fees: Costs for any external consultants or experts brought in for specialized tasks
Training Costs: Expenses for training staff and preparing training materials.
Data Migration Costs: Costs involved in migrating existing data to the new system, including data cleansing and validation.
Miscellaneous Costs: Other costs such as marketing for the new system, administrative expenses, and contingency funds.

2.2.2 Operational Savings

Efficiency Gains: Reduction in time and resources spent on manual processes and data management.
Error Reduction Savings: Decrease in costs associated with data errors, such as incorrect policy quotes or claims processing mistakes.
Staff Productivity: Improved productivity due to streamlined processes and automation.
Maintenance Savings: Savings from reduced need for maintaining old systems or manual processes.

2.3 Detailed business case of the application

The proposed project involves the development of a comprehensive database and interactive system tailored for an insurance company. This initiative is expected to enhance operational efficiency, improve customer service, ensure compliance, and drive overall business growth.
The insurance sector is increasingly reliant on technology to manage data, interact with customers, and streamline processes. To develop a robust system that addresses these needs, ensuring better data management, customer interaction, and operational efficiency.

3 Data description and visualization

3.1 Data description

The dataset used in this project is downloaded from Kaggle, the world's largest data science community with powerful tools and resources to help developer achieve their data science goals. It contains 986 entries with 11 columns. The features and the data types are as followed:

Age: Integer
Diabetes: Integer
BloodPressureProblems: Integer
AnyTransplants: Integer
AnyChronicDiseases: Integer
Height: Integer
Weight: Integer
KnownAllergies: Integer
HistoryOfCancerInFamily: Integer
NumberOfMajorSurgeries: Integer
PremiumPrice: Integer

The average age in the dataset is around 41.4 years, with a standard deviation of 13.7 years. The average height is approximately 168 cm and the average weight is around 75.9 kg. There are no missing values in any of the columns.

3.2 Data visualization

4 Features analysis

4.1 Heat map

Visualization of the optimal solutions plays a very important role in multi-objective optimization (MO). In MO with conflicting objectives there is no single optimum, and search methods return a set of solutions from which one must be selected. In order to select the solution, a decision maker usually needs to visualize the discovered solutions in the objective space. Heatmap visualization is a technique which is most often applied to data gathered from micro-arrays.[1] In the heat map in our project, the warmer color representing a stronger positive relationship. The strongest correlation between the column are between the feature age and premium price.

4.2 XGboost

Machine learning, based on statistics and big data, is quickly emerging to reduce the development time of new and improved materials, and thereby promote the progress of big data in the material industry.[2] This importance score chart presents an analysis of the relative importance of various health-related features in a predictive model used for assessing health risks. The model considers factors such as age, medical history, and lifestyle attributes. By assigning importance scores to each feature, the analysis helps in understanding which factors are more influential in predicting health risks.

Age (Importance Score: 0.35) Age is a critical determinant of health risk, with older individuals generally at higher risk for various health conditions. Focus on age-specific screening and preventive measures.

Diabetes (0.02) While diabetes is a significant health concern, its relative importance in this model is lower, possibly due to its manageable nature with medication and lifestyle changes. Routine monitoring and lifestyle advice for diabetic patients.

Blood Pressure Problems (0.03) Blood pressure issues, though relatively low in importance, can lead to serious complications if not managed. Regular monitoring and lifestyle modifications.

Any Transplants (0.31) Transplant history is a major indicator of health risk due to the associated long-term medication and potential complications. Close monitoring of transplant patients for any signs of complications.

Any Chronic Diseases (0.11) The presence of chronic diseases significantly impacts overall health risk. Comprehensive management plans for chronic conditions.

Height (0.04) Height has a moderate impact, potentially due to its association with certain health risks. Correlational analysis with specific health conditions.

Weight (0.06) Weight, particularly obesity, is a known risk factor for many health issues. Weight management programs and dietary counseling.

Known Allergies (0.01) Allergies have the least impact on overall health risk in this model. Standard allergy management and education.

History of Cancer in Family (0.05) A family history of cancer moderately influences health risk assessments. Regular screenings for individuals with a family history of cancer.

Number of Major Surgeries (0.07) Multiple major surgeries indicate a higher health risk, possibly due to underlying health conditions or the impact of surgeries themselves. Post-surgery follow-ups and comprehensive health assessments.

5 Environment Configuration and Dependencies

Our app is based on Python 3.9 and SQL 8.0. We have deployed the SQL 8.0 database on Microsoft Cloud, and it uses SSL to connect to the Microsoft Cloud database. We provide the SSL CA file and a config file. To connect to the cloud, You only need to modify the "input_file_path" in the "get_config_dict" function within our "end2endapp.py" to load the config file, and configure the path of the ssl_ca file in "config.yaml". Besides, the 'xgb_reg.pkl' file path also need to be modified.
The Python dependencies include scikit-learn, xgboost, pandas, seaborn, PyYAML, easydict, Flask, and mysql-connector-python.

6 How to run our application

After completing the environment setup, execute python end2endapp.py, and then use a browser to visit http://127.0.0.1:5000 to start our application. By default, it connects to our database, but you can also modify the config.yaml file to connect to your own database. We provide a finalProjectDatabase.sql file to create all the tables in an empty database. It implements the EER schema from part 2 as an SQL relational database.

7 Implementation and APIs

7.1 Front-end

We have implemented the web front-end interaction design using the Python-Flask framework, allowing both potential customers and company employees to access the content they wish to view.
On our website, company administrators can select to view database table data, add new content, or delete and update table contents by typing in the row number.
For users interested in purchasing insurance, we provide an online questionnaire

to survey their health status. The insurance price is then calculated by calling our machine learning algorithm.

7.2 Back-end

We have deployed a SQL 8.0 server on Microsoft Azure Cloud and set up the finalProject database on it. The creation of database tables and their subsequent data population were completed using "finalProjectDatabase.sql". In "access_database.py", we provide two methods. "print_table_info" is used to display information about the database and all its tables, while "operating_mydatabase" offers an interface for database administrators to insert, delete, and update data.

7.3 Machine Learning part

Our machine learning model's training data and methods have already been elucidated in part 3. Hence, here we simply invoke the trained model to implement the function of insurance price assessment. If there is new data available, you can use the methods described in part 3 to retrain our model.

8 Simple Tutorial for Our Website

8.1 User Guide

On the main page of the website, click 'Get Started!' and follow the instructions to fill out the questionnaire. You will receive a customized insurance price.
Fill in the questionnaire:

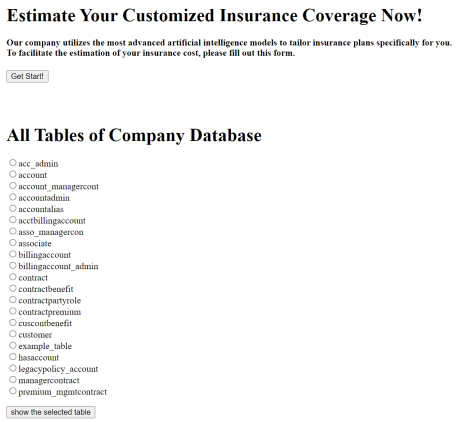

Figure 1: click 'Get Start!'

Please fill in this form

Age:
22
Have you ever suffered from diabetes (0 No, 1 Yes):
0
Have you ever suffered from any blood pressure-related illnesses? (0 No, 1 Yes):
1
Have you ever suffered from any transplants (0 No, 1 Yes):
1
Have you ever suffered from any transplants chronic diseases (0 No, 1 Yes):
0
Height:
170
Weight:
75
Do you have any known allergic reactions? (0 No, 1 Yes):
0
Does your family have a history of cancer? (0 No, 1 Yes):
0
How many major surgeries have you undergone in the past?
1

submit

Prefer to fill in later?

Go back

Figure 2: fill in the questionnaire

Get your price:

Your estimated price is:

Age	22
Diabetes	0
any blood pressure-related illnesses?	1
any Transplants?	1
any transplants chronic diseases?	0
Height	170
Weight	75
any known allergic reactions	0
history of cancer?	0
number of major surgeries?	1

Your estimated insurance price is: 35472.285

Try another?

Go back

Figure 3: get your price

8.2 Administrator Guide

On the main page of the website, you will see all the current tables in the database. You can access the table information by selecting a table and clicking the 'show the selected table' button. This allows you to view all the data in the selected table.

8.2.1 Insert

To insert new data, you need to select the 'insert data' option on the page where you are viewing table information and click the 'submit' button. After that, you will see the system prompting you to enter the value for each column of the new data. Follow the instructions (please ensure valid input), and finally, click the 'Insert' button to confirm submission. Upon returning to the table viewing interface, you will find that the insertion has been successful.

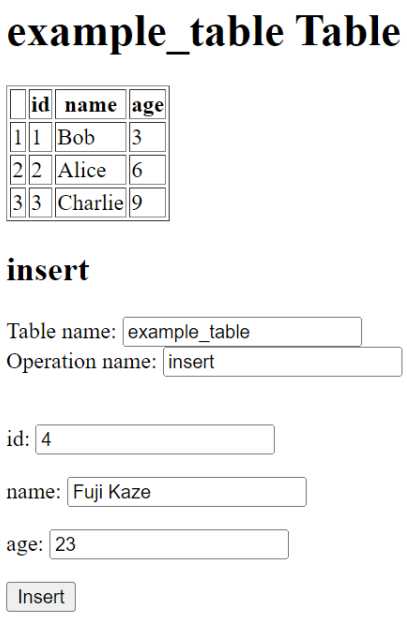

Figure 4: enter values

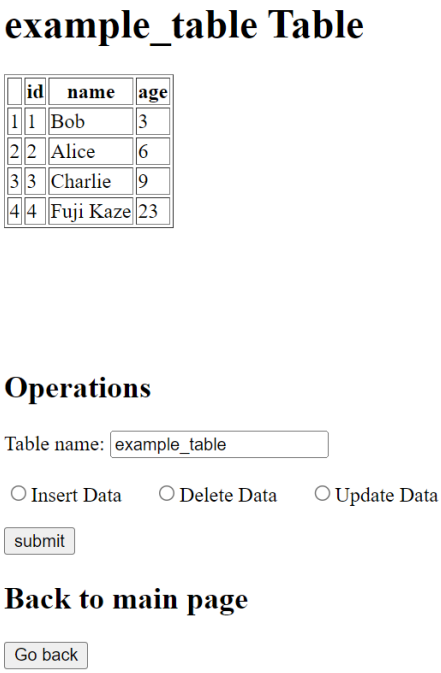

Figure 5: insert success

8.2.2 Delete

To delete data, you need to select the 'delete data' option on the page where you are viewing table information and click the 'submit' button. After that, you will see the system prompting you to enter the value for row number that you want to delete. Follow the instructions (please ensure valid input), and finally, click the 'Delete' button to confirm submission. Upon returning to the table viewing interface, you will find that the delete has been successful.

example_table Table

	id	name	age
1	1	Bob	3
2	2	Alice	6
3	3	Charlie	9
4	4	Fuji Kaze	23

delete

Table name: example_table

Operation name: delete

Delete row number: 3

Delete

Figure 6: delete row

example_table Table

	id	name	age
1	1	Bob	3
2	2	Alice	6
3	4	Fuji Kaze	23

Operations

Table name: example_table

○ Insert Data ○ Delete Data ○ Update Data

submit

Back to main page

Go back

Figure 7: delete success

8.2.3 Update

To update data, you need to select the 'update data' option on the page where you are viewing table information and click the 'submit' button. After that, you will see the system prompting you to enter the value for row number that you want to update and the value for each column of the new data. Follow the instructions (please ensure valid input), and finally, click the 'Delete' button to confirm submission. Upon returning to the table viewing interface, you will find that the updatehas been successful.

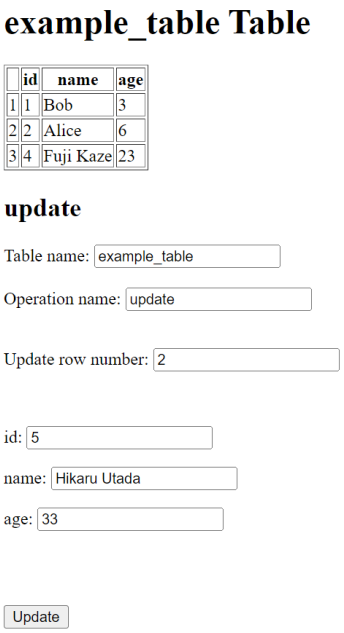

Figure 8: update row

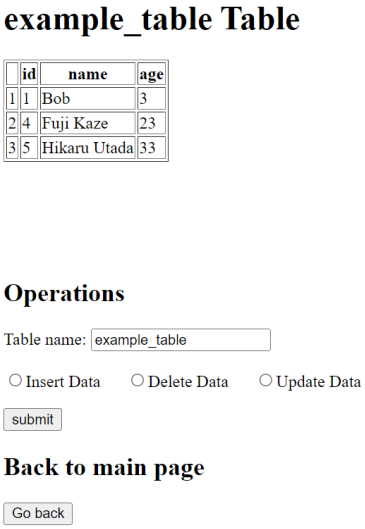

Figure 9: update success

9 Conclusion

In this final project, we have implemented the EER diagram from parts 1 and 2 through a cloud-based SQL database. Moreover, we trained a machine learning model using a Kaggle dataset to fulfill the customized insurance needs of both the company and its clients in Part3. Finally, we utilized a web-based app to provide end-to-end quick access, allowing both users and company management to swiftly meet their respective needs through web access.

10 Reference

1. Pryke, A., Mostaghim, S., Nazemi, A. (2007). Heatmap Visualization of Population Based Multi Objective Algorithms. In: Obayashi, S., Deb, K., Poloni, C., Hiroyasu, T., Murata, T. (eds) Evolutionary Multi-Criterion Optimization. EMO 2007. Lecture Notes in Computer Science, vol 4403. Springer, Berlin, Heidelberg. https://doi.org/10.1007/978-3-540-70928-2-29.
2. Xiupeng Shi, Yiik Diew Wong, Michael Zhi-Feng Li, Chandrasekar Palanisamy, Chen Chai, A feature learning approach based on XGBoost for driving assessment and risk prediction, Accident Analysis Prevention, Volume 129, 2019, Pages 170-179, ISSN 0001-4575, https://doi.org/10.1016/j.aap.2019.05.005.

11 Github Links

CSCI2433 2023 fall project part 1-3
CSCI2433 2023 fall project