Collaborators: Jiayu Gu, Denghao Sun

Date: December 7, 2023

NetID: jg7956, ds6963

Section: 001

**Database Systems Project Part III**

**Logical Schema Optimization and Machine Learning Model Creation**

**Affirmation of our independent efforts: \_\_\_\_Denghao Sun, Jiayu Gu\_\_\_\_\_\_\_\_\_\_\_**

Machine Learning for forecast chronic disease( I use excel and rapidMiner to do the Machine learning):

* To predict the risk of chronic diseases for use by an insurance company, we need to gather and analyze customer information. Our approach involves creating a table for estimating the risk of chronic diseases among customers. The table includes the following columns:
  + **Customer's Place of Residence:** We analyze the address based on the city. People living in urban areas have a higher likelihood of developing chronic diseases. Therefore, if the customer's city of residence is classified as an urban area, the 'Urbanization' column is assigned a value of 1; otherwise, it is set to 0.
  + **Customer's Income:** We categorize family income into three groups: low income (assigned a value of 2), middle income (assigned a value of 1), and high income (assigned a value of 0).
  + **Age and Gender of Customers:** Research suggests that men are more susceptible to major life-threatening chronic diseases. Thus, if the customer's gender is male, the 'Gender' column is set to 1; otherwise, it is set to 0. Additionally, if the customer's age is greater than 50, the 'Age' column is set to 1; otherwise, it is set to 0.
  + **Genetic Information:** We use genetic information to determine if the customer has a genetic predisposition to chronic diseases. If there is a hereditary link to chronic diseases, the 'Heredity of Chronic Disease' column is assigned a value of 1; otherwise, it is set to 0.
  + **Boolean Columns Factors**:
    - Insufficient physical exercise=1, sufficient physical exercise=0
    - Unhealthy eating habit=1, healthy eating habit=0
    - Smoking and drinking=1, no Smoking and drinking=0
    - Unstable emotion status=1, stable emotion status=0
    - Burning the midnight oil=1, no Burn the midnight oil=0
* **Scoring and Risk Classification:** Finally, we introduce a new column where we sum up all the scores. If the total score is greater than or equal to 6, the individual is classified as 'High Risk.' Conversely, if the total score is less than 6, the classification is 'Low Risk.'

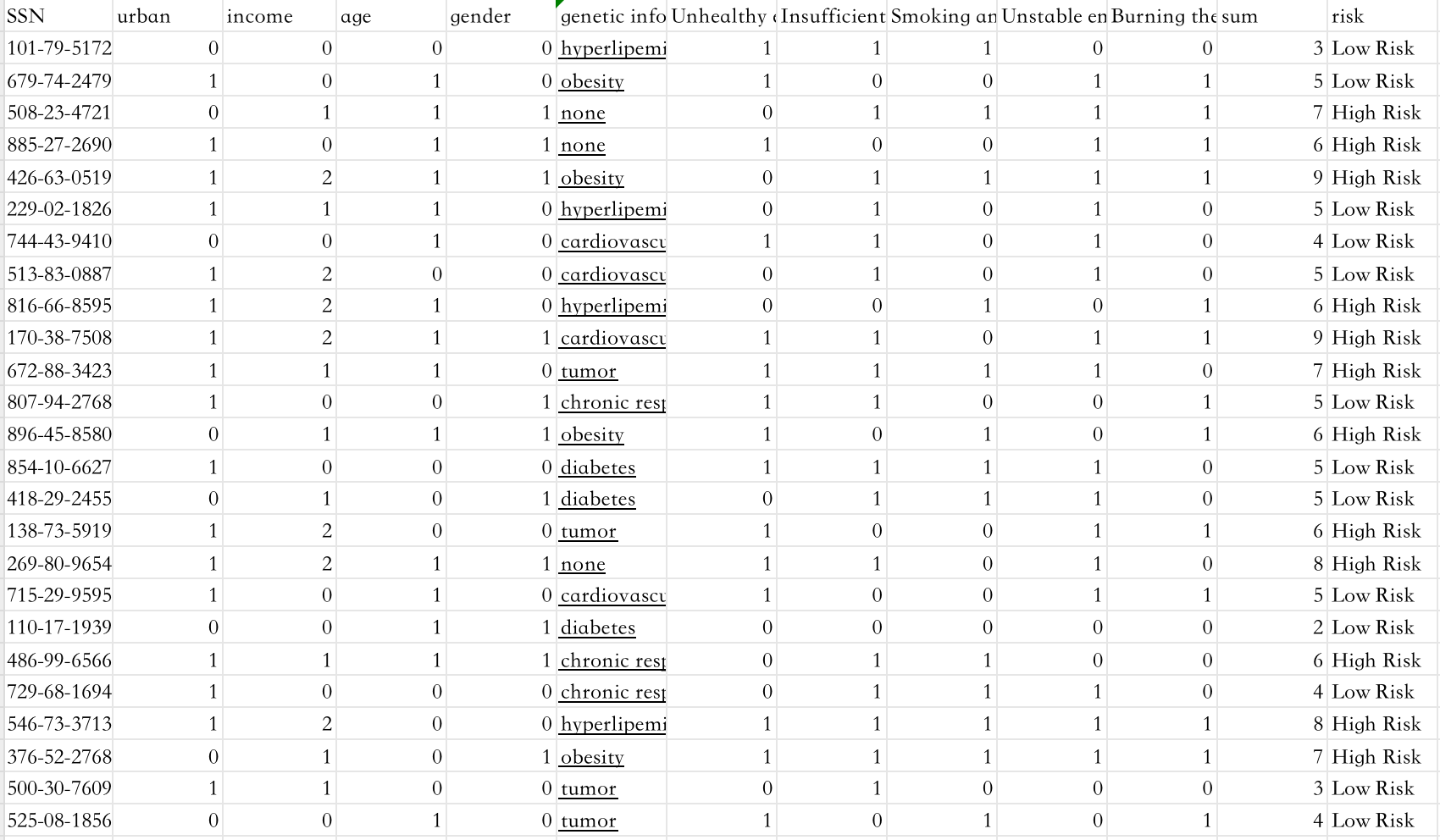
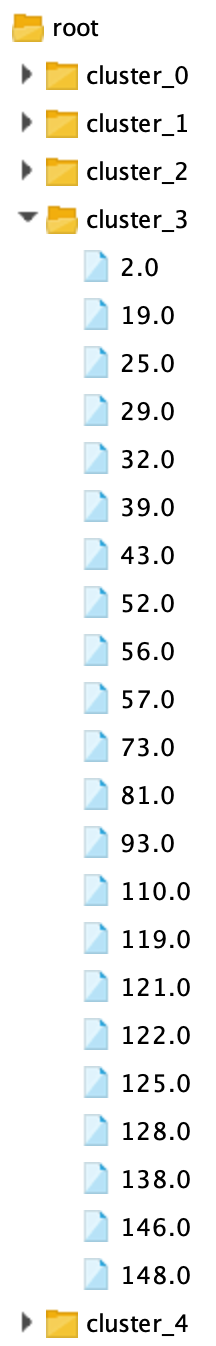
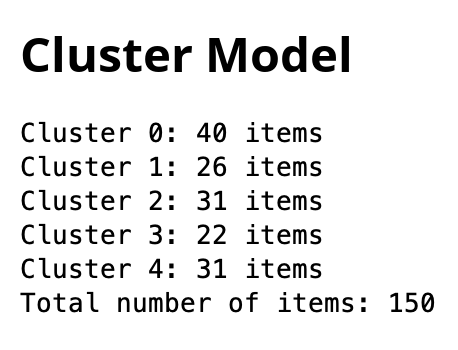
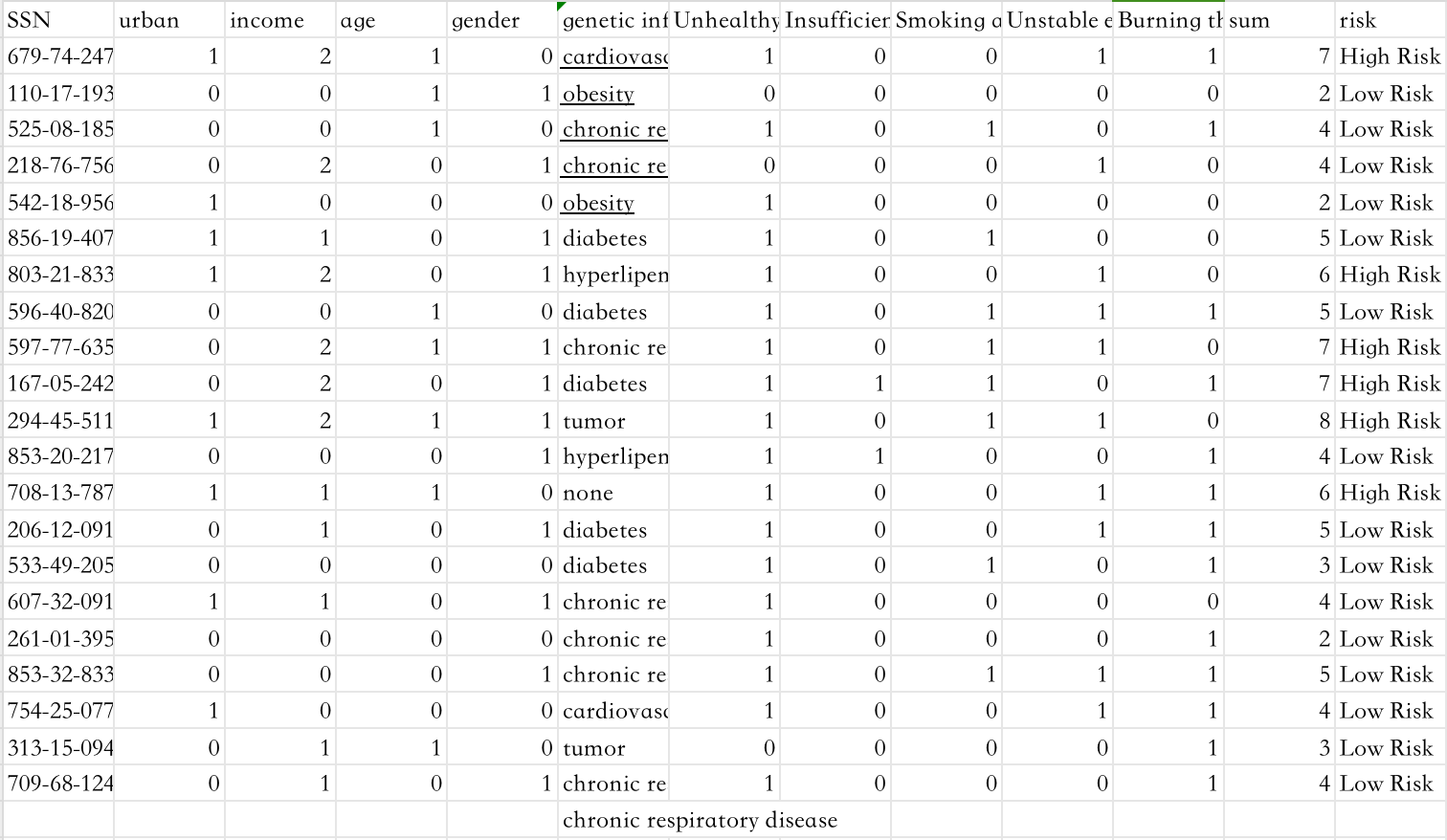


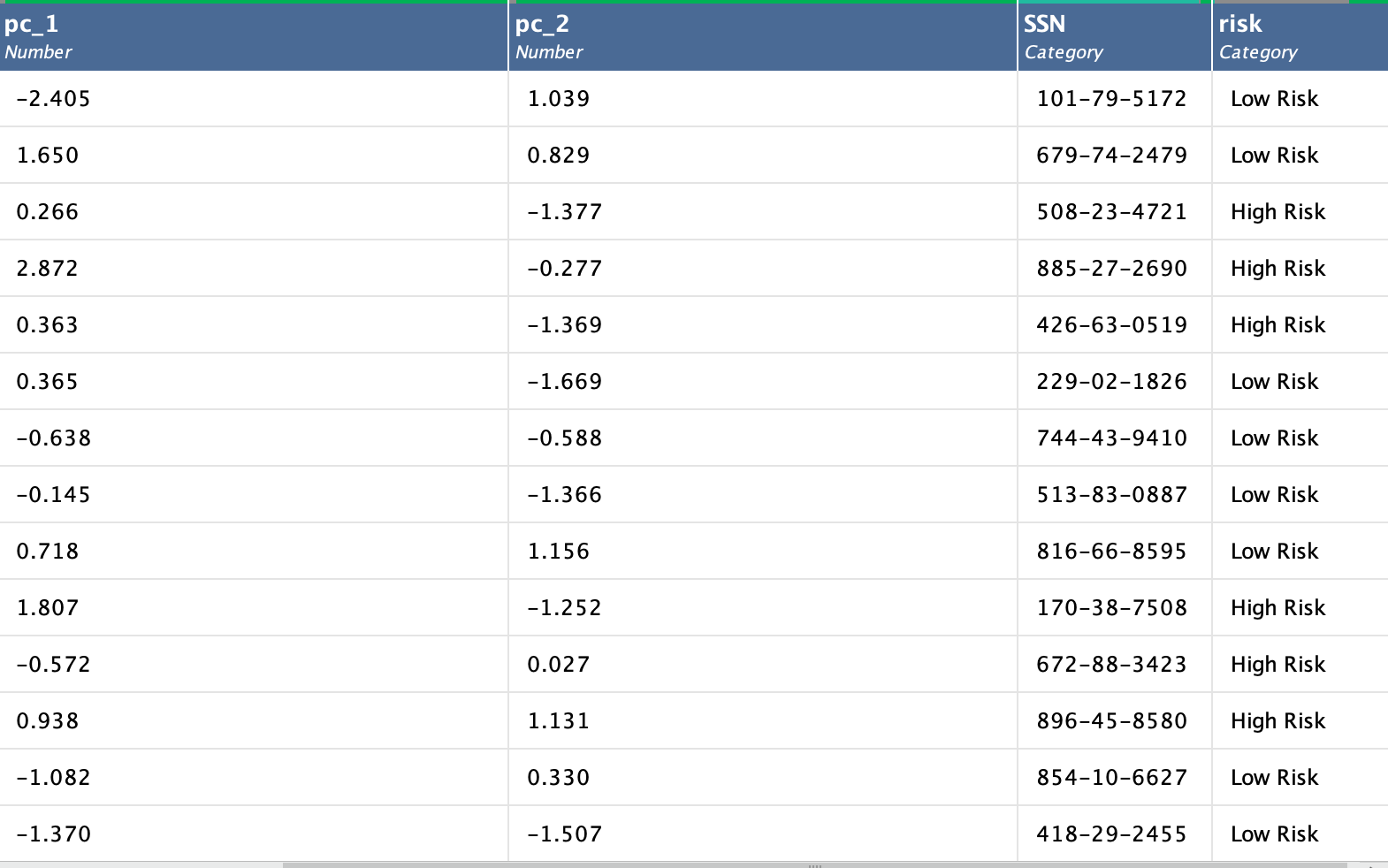
Figure 1 the table of customers with factors

* **K-Mean And Cluster:** We can divide the customer into different groups based on their information. We use K-means to divide into different groups. Then, for each group, we can get a max frequency of a chronic disease , like cardiovascular disease and diabetes based on some related disease like genetic information or the medical record if they have. Finally, we divide all customers into

Figure 2 the result of k-mean Figure 3 one group of cluster Figure 4 the information for one cluster and the name of chronic disease in last row

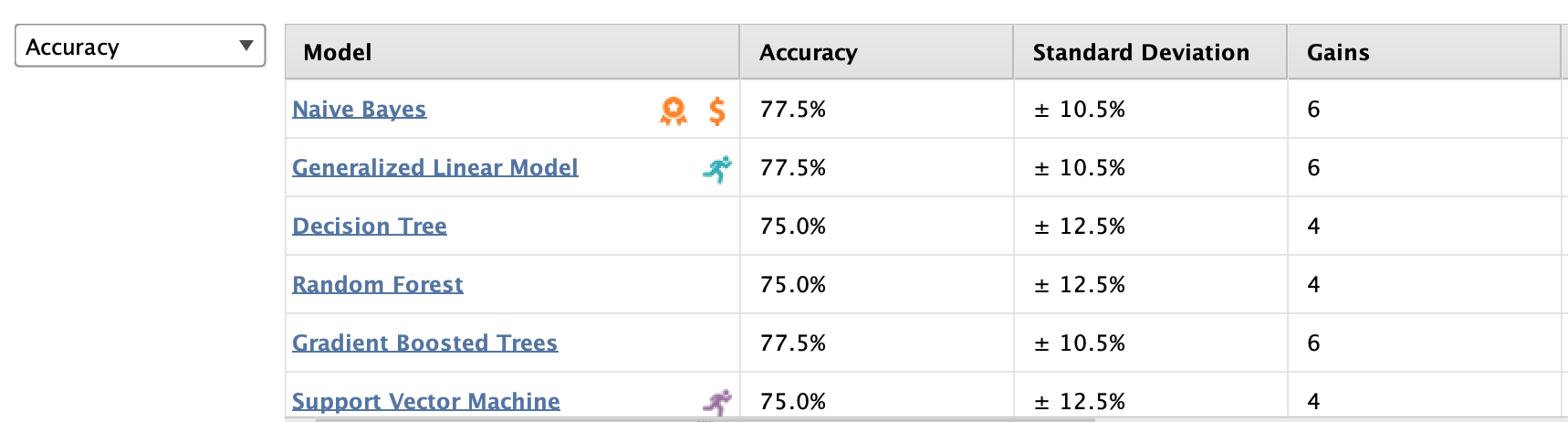
different groups with specific chronic disease. Each customer in the cluster will be estimated as high risk or low risk. When there is a new customer add, he will automic into a specific cluster.

* **Reduce Dimension:** We normalize and outlier the data in each

Figure 5 the table after normalization, outlier and PCA

and use PCA to reduce the data dimension to improve the accuracy of prediction.

* **Modeling:** Then we choose an appropriate model like a decision tree or random forest to do the prediction for each chronic disease group. The result of the model will test if the factors are important for the result and the accuracy of prediction.

Figure 6 The different models for one cluster

* **Product Development:** Based on the distinct customer groups associated with specific chronic diseases, the insurance company can develop tailored insurance products and policies. These products may focus on particular health-related policy options.
* **Assessment Risk:** If the insurance product is about cardiovascular disease, we can go to the cluster about cardiovascular disease. There are some factors that can cause cardiovascular disease. Therefore, we can use ML to find which are some important factors that will cause cardiovascular disease. Then use the same method above to assess the risk of disease.
* **Health Monitoring and Alerts:** Collaborating with health monitoring companies, the insurance company can continuously track customer health. If a customer's health status changes, the system will automatically adjust their risk level. Persistent high-risk status may trigger alarms, leading to potential adjustments in insurance costs.
* Short term benifits:
  + **Cost Reduction:** By categorizing customers into high-risk and low-risk groups, the insurance company can offer different policies with varying costs, optimizing resource allocation.
  + **Competitive Advantage:** Utilizing predictive models allows the insurance company to provide more personalized and competitively priced policies, enhancing its market position.
  + **Risk Mitigation:** Early alerts enable high-risk individuals to take proactive measures to manage and mitigate their health risks.
* Medium term benefits:
  + **Data Accuracy:** With an increasing number of customers using insurance products, the quality of data improves, leading to more accurate predictions.
  + **Product Innovation:** Clustering customer information reveals shared characteristics, facilitating the creation of innovative insurance products.
  + **Customer Retention:** As customers experience the benefits of early disease detection and preventive measures, loyalty to the insurance company may increase.
* Long Medium term benefits:
  + **Profitability Increase:** Early alerts reduce treatment costs for customers and limit the number of insurance claims, resulting in increased profitability.
  + **Reduced Chronic Diseases:** Over time, proactive risk management can contribute to a reduction in the overall prevalence of chronic diseases in society.

**EDA Physical Design Part:**

**Actions Taken**: The physical database for the application has been deployed in MySQL. This involved establishing a set of tables and defining indexes and key constraints to facilitate efficient data storage, retrieval, and integrity within our system's architecture.

**Indexing Key Columns in several Tables**

Queries like:

ALTER TABLE Claim ADD INDEX (SSN);

ALTER TABLE Claim ADD INDEX (LineOfBusiness);

ALTER TABLE Claim ADD INDEX (ContractNumber);

Rationale: Indexing these columns in the Claim table enhances the performance of queries filtering by SSN, LineOfBusiness, and ContractNumber. These fields are likely used in join operations and where clauses, making indexes on them beneficial for query speed.

Indexing Key Columns in CustomerAlias Table

**Partitioning the BillingAddress Table**

Query: PARTITION BY HASH (AddressID) PARTITIONS 4;

Rationale: Implementing hash partitioning on the AddressID column of the BillingAddress table helps distribute the data more evenly across partitions. This can enhance query performance, especially in a large dataset, by reducing the amount of data scanned per query.

Composite Indexing on Various Tables

**Composite Indexing on Various Tables**

Queries:

CREATE INDEX idx\_invoice\_cusssn\_duedate ON Invoice (CusSSN, DueDate);

CREATE INDEX idx\_accountproduct\_accountid\_productline ON Account\_Product (AccountID, ProductLineOfBusiness);

CREATE INDEX idx\_account\_accountid\_taxid ON Account (AccountID, TaxIDNumber);

CREATE INDEX idx\_customers\_ssn\_lastname ON Customers (SSN, CusLastName);

CREATE INDEX idx\_customers\_firstname\_lastname\_dob ON Customers (CusFirstName, CusLastName, CusDOB);

CREATE INDEX idx\_payment\_creditcardno\_bankingaccount ON Payment (CreditCardNo, BankingAccountNumber);

CREATE INDEX idx\_payment\_cardtype\_expirationdate ON Payment (CardType, ExpirationDate);

CREATE INDEX idx\_contract\_ssn\_billingmethod ON Contract (SSN, BillingMethod);

CREATE INDEX idx\_employee\_lastname\_firstname\_state ON Employee (EmpLastName, EmpFirstName, EmpState);

CREATE INDEX idx\_product\_lineofbusiness\_seriesname ON Product (LineOfBusiness, SeriesName);

CREATE INDEX idx\_contractpremium\_premiumcode\_processdate ON ContractPremium (PremiumCode, ProcessDate);

Rationale: Composite indexes were created on columns frequently used together in queries. This approach is particularly effective for optimizing queries that filter or join tables based on these column combinations, leading to faster query execution and reduced load times.

**Business use cases:**

* Use Case 1: Customer Registration
  + Process: The customer accesses the registration interface, submits their details, and receives a confirmation upon successful registration.
* Use Case 2: Quote Request
  + Process: The customer inputs the necessary data required for obtaining a quote, which the system processes to provide a preliminary quote.
* Use Case 3: Policy Customization and Application
  + Process: The customer reviews, customizes, and submits their policy preferences. The system records the application and initiates the underwriting process.
* Use Case 4: Underwriting and Policy Issuance
  + Process: The underwriting department evaluates the application, and upon approval, the system issues the policy and notifies the customer.
* Use Case 5: Payment Setup
  + Process: The customer sets up their preferred payment method, and the system schedules future payments and generates invoices.
* Use Case 6: Claims Filing and Processing
  + Process: The customer files a claim when needed, and the system manages the claim processing, including document management and settlement.