Healthcare Insurance Data Analysis Using Big Data Technologies

A Project Report

Authors:

Felix Luca Krebs (2470475)
MD Kamruzzaman Russel (2470478)
Justus Izuchukwu Onuh (2470477)

Ho Chi Minh City University of Technology, HCMUT

April 26, 2025

1 Introduction

Healthcare insurance companies today face significant challenges in understanding customer behavior, forecasting future risks, and maximizing revenue in an increasingly competitive environment. Traditional methods of data analysis are no longer sufficient due to the explosive growth of healthcare-related data generated from various sources including hospitals, insurance claims, subscriber activities, and public health records.

To address these challenges, a healthcare insurance company decided to leverage Big Data technologies to analyze large datasets collected from multiple heterogeneous sources — including hospital patient records, subscriber information, and disease reporting systems. By applying advanced analytics on this data, the company aims to better understand customer needs, predict potential health risks, customize offers, manage insurance claims efficiently, and ultimately reward loyal customers who contribute to the organization's growth.

Big Data systems, particularly tools such as Apache Hadoop, Apache Spark, and Apache Hive, are uniquely suited for this task because they provide fast, scalable, distributed storage and processing capabilities. Hadoop enables the storage of massive datasets across clusters of machines, Spark offers in-memory processing for high-speed analytics, and Hive allows querying structured data using SQL-like syntax.

In this project, we focus on analyzing healthcare claims, patient demographics, and disease patterns to extract meaningful insights that can guide business strategies. The objectives of this project include:

- Cleaning and preparing healthcare data for analysis.
- Exploring and visualizing trends related to insurance claims and patient behavior.
- Designing sample Big Data workflows using Spark and Hive for future large-scale deployments.

This work references and is inspired by the GitHub repository HELTHCARE-SYSTEM, which provides a practical example of how Big Data ecosystems can be effectively utilized in healthcare insurance systems.

2 Project Structure and Methodology

This project was structured according to a Big Data analytics pipeline, encompassing the key stages of data ingestion, cleaning, processing, merging, and visualization. Each stage was carefully designed to ensure that meaningful insights could be derived from the raw healthcare data.

2.1 Data Collection

The first phase involved collecting multiple datasets from different sources. These datasets included:

- **Patient Records:** Containing demographic information, gender, birth dates, and disease history.

- **Insurance Claims:** Containing details of insurance claims filed by patients, including claim amounts, types, and outcomes.
- **Hospital Information:** Detailing the hospitals where patients were treated.
- **Subscriber Data:** Providing subscriber identifiers and linking subscribers to claims and healthcare services.

2.2 Data Cleaning

Before analysis, the raw data underwent an intensive cleaning process:

- Duplicates were identified and removed.
- Missing values were handled appropriately, either by removal or imputation depending on the column relevance.
- Data types were standardized, with dates converted into proper datetime formats and numeric fields ensured for computations.
- Column names were normalized to maintain consistency across different datasets.

2.3 Data Merging

Following cleaning, the datasets were integrated:

- Claims were linked to patient demographic data using the patient identifiers (patient_id and Patient_id).
- Subscriber information was merged using subscriber identifiers (SUB_ID) to track and analyze customer behavior.
- Hospital information was associated with claims data to analyze patterns in claim distribution among hospitals.

2.4 Data Exploration and Visualization

The merged datasets were subjected to exploratory data analysis (EDA) to extract patterns and trends. Key visualizations included:

- The distribution of claim amounts across the population.
- Identification of the top 10 diseases based on claim frequencies.
- Analysis of hospitals with the highest average claim amounts.
- Investigation of relationships between patient age and claim behavior.
- Profiling of subscribers who filed the most claims.

The visualizations provided actionable insights into customer behavior, hospital practices, and disease trends.

2.5 Technology Stack

- **Data Processing and Analysis:** Python 3, Pandas for data manipulation.
- **Visualization:** Seaborn and Matplotlib for creating statistical charts and visual representations.
- **Notebook Environment:** Jupyter Notebooks were utilized for iterative analysis and visualization.
- **Big Data Technologies (Optional Extension):**
  - **Apache Spark:** A sample PySpark script (customer_behavior_analysis.py) was provided to demonstrate how large datasets could be processed in a distributed manner.
  - **Apache Hive:** A sample HiveQL script (claim_summary.hql) was included to showcase how healthcare data could be queried using SQL-like syntax in a Big Data environment.

The use of Spark and Hive would be essential for scaling this solution to enterprise-level data volumes.

3 Experiments

After preparing and cleaning the datasets, several exploratory analyses were conducted to uncover key patterns, trends, and business insights from the healthcare insurance data.

3.1 Distribution of Claim Amounts

The distribution of insurance claim amounts was analyzed to understand the overall financial risk landscape. The histogram revealed that the majority of claim amounts were relatively low, clustered towards the lower end of the scale. However, a few instances of very high claims were observed, acting as outliers. This skewed distribution is typical in insurance contexts, where many small claims are common, but large claims, although rare, have a significant impact on total payouts. Identifying and managing such outliers is crucial for optimizing insurance reserves and risk modeling.

3.2 Top 10 Subscribers by Number of Claims

An analysis of claim frequencies per subscriber highlighted that a small subset of subscribers were responsible for a disproportionately large number of claims. These high-frequency claimants could potentially pose a higher financial risk to the company. Recognizing these patterns enables the insurance company to implement proactive strategies, such as closer monitoring, specialized policy designs, or differentiated pricing models, targeted towards high-risk customer segments.

3.3 Top 10 Hospitals by Average Claim Amount

By linking hospital data with claims, we identified the top hospitals associated with the highest average claim amounts. Certain hospitals consistently exhibited higher billing practices compared to others. This finding suggests the need for targeted audits, renegotiations of service agreements, or strategic partnerships with hospitals to control costs and ensure billing transparency.

3.4 Top 10 Reported Diseases

The frequency distribution of reported diseases revealed the most common medical conditions leading to insurance claims. Understanding which diseases are most prevalent allows the insurance company to tailor its product offerings, prepare actuarial risk models, and develop health management programs that align with actual customer needs. It also highlights potential areas where preventive care initiatives could be promoted.

3.5 Claim Amount vs. Age of Patients

A scatterplot analysis examining the relationship between patients' ages and their claim amounts indicated that younger patients tended to have smaller claims compared to older patients. This trend aligns with general health risk profiles, where older individuals are statistically more likely to incur higher healthcare costs. Such insights are vital for designing age-based premium structures and refining underwriting policies.

Each of these exploratory experiments contributed critical business intelligence, offering actionable strategies for customer segmentation, risk management, hospital network optimization, and policy design. Collectively, they demonstrate the value of Big Data analytics in enhancing decision-making within the healthcare insurance sector.

4 Improvements or Application Proposal

While the current project successfully demonstrates the potential of Big Data analytics in the healthcare insurance sector, several improvements and future applications could significantly enhance the system's effectiveness and strategic value.

4.1 Predictive Modeling

The healthcare insurance company could develop machine learning models to predict customer behaviors, specifically focusing on identifying individuals who are at higher risk of submitting large insurance claims. Techniques such as logistic regression, decision trees, or ensemble methods could be employed to classify subscribers based on risk levels. Proactive risk management strategies could then be implemented, such as targeted premium adjustments, special wellness programs, or early intervention offers, thereby optimizing financial reserves and improving customer satisfaction.

4.2 Real-Time Analytics

Currently, the analysis operates in a batch mode, analyzing historical datasets. Transitioning to a real-time analytics system would provide substantial advantages. By integrating technologies such as Apache Kafka for real-time data ingestion and Apache Spark Streaming for live data processing, the company could monitor claims submissions, hospital billing patterns, and customer interactions as they occur. This would enable faster fraud detection, real-time customer engagement, and dynamic policy adjustments, greatly enhancing operational agility.

4.3 Customer Segmentation

Implementing clustering algorithms, such as K-Means, could allow the company to segment its subscriber base into distinct groups based on claim behavior, demographics, healthcare usage, and spending patterns. These customer segments could then be targeted with personalized insurance products, marketing campaigns, and wellness initiatives, maximizing revenue while improving customer loyalty and satisfaction.

4.4 Data Enrichment

The predictive power and depth of analysis could be further improved by integrating external datasets, including:

- **Regional Health Statistics:** Linking customer data with public health trends to anticipate disease outbreaks.
- **Socio-economic Data:** Understanding how income, education, and living conditions affect claim behaviors.
- **Environmental and Weather Data:** Identifying patterns where environmental factors might increase certain health risks (e.g., asthma during high pollution periods).

Such enriched datasets would provide a holistic view of customer risk profiles and enable the development of more accurate predictive models and strategic initiatives.

5 Conclusion

In this project, we demonstrated the power and relevance of Big Data analytics in addressing the challenges faced by healthcare insurance companies. By systematically collecting, cleaning, merging, and analyzing healthcare datasets, we were able to extract meaningful insights regarding customer behavior, hospital billing patterns, and disease trends.

The exploratory analysis revealed key patterns, such as the concentration of claims among a small group of subscribers, the disproportionate billing practices among certain hospitals, and the relationship between patient age and healthcare costs. These insights provide a strong foundation for more strategic business decisions, including targeted risk management, customer segmentation, and product personalization.

Although this project primarily operated on small-to-medium datasets using tools like Python and Jupyter Notebooks, it was carefully structured to be scalable. We provided sample Spark and Hive scripts to illustrate how the same workflows could be expanded to handle enterprise-scale Big Data environments, ensuring real-world applicability.

Future enhancements, such as predictive modeling, real-time analytics, and data enrichment, offer exciting opportunities to further optimize insurance operations and enhance customer experiences.

This project highlights the transformative potential of Big Data technologies in the healthcare insurance sector, and lays a blueprint for intelligent, data-driven decision making.

6 References

- Tejas Bansal, "Big Data Analytics for Healthcare Systems", GitHub Repository, https://github.com/tejasjbansal/HELTHCARE-SYSTEM.
- Official Pandas Documentation, https://pandas.pydata.org/docs/
- Seaborn Visualization Library, https://seaborn.pydata.org/