International Journal of Emerging Trends in Computer Science and Information Technology
ISSN: 3050-9246 | https://doi.org/10.63282/3050-9246.IJETCSIT-V3I2P105
Eureka Vision Publication | Volume 3, Issue 2, 42-52, 2022

Original Article

Predicting Eligibility Gaps in CHIP Using BigQuery ML and Snowflake External Functions

Parth Jani
IT Project Manager at Molina HealthCare, USA.

Abstract - *Using SQL-based ML technologies specifically, Google Big Query ML & Snowflake External Functions this study explores the latest approach for identifying & projecting their eligibility gaps in the Children's Health Insurance Program (CHIP). The primary goal is to enhance their public health projections by means of their proactive identification of those at risk of losing CHIP coverage resulting from administrative mistakes, changeable income, or inadequate documentation. By integrating demographic data and structured healthcare into Big Query, one may immediately train machine learning models within SQL environments, hence removing traditional data engineering limitations and accelerating model deployment. Concurrent with this, Snowflake External Functions enabled simple access to third-party APIs and cloud services, hence improving contextual insights and supporting dynamic rule application. By means of their combined usage, these systems provide a scalable and affordable method to expose trends and risk indicators often hidden within large-scale statistics. Our results suggest that this paradigm might effectively predict potential eligibility interruptions, hence allowing more timely interventions and legislative changes. The study emphasizes the increasing importance of SQL-based ML technology in public sector projects, especially in situations where time-sensitive decisions influence their vulnerable groups. By allowing data analysts to work within familiar environments, these technologies democratize their access to advanced analytics & thereby support fast & informed decision-making in healthcare systems. This work argues for data-native, ML-driven approaches in public health management, therefore improving a proactive, data-informed model of care continuity.*

Keywords - *CHIP, eligibility gaps, BigQuery ML, Snowflake, external functions, SQL-based ML, gap-filling models, predictive analytics, healthcare lapses, Medicaid churn.*

1. Introduction

Since its founding in 1997, the Children's Health Insurance Program (CHIP) has been a cornerstone of public health in the United States, offering millions of low-income children ineligible for their Medicaid necessary medical care. Maintaining continuous coverage under CHIP is not just for the health & wellbeing of children but also for limiting long-term healthcare expenses & therefore reducing the burden on emergency medical services. Children who have constant access to healthcare are more likely to undergo frequent tests, vaccines & early treatments meant to prevent their chronic diseases or disorders. Still, coverage gaps remain a major obstacle, mostly resulting from administrative & more systematic inefficiencies rather than actual changes in eligibility. This event, often known as administrative turnover, can cause children either temporarily or permanently to lose access to more essential medical treatment.

Administrative turnover results from various elements: inadequate paperwork, missing deadlines, changing family finances, or more inadequate government communication. These shortcomings especially worry me as they disproportionately affect impoverished households with limited resources or internet connection. Many of these children eventually re-enrolled, but the break in care might cause delayed their treatment and poorer medical results. Even little pauses might discourage families from reapplying, cause financial difficulty from unpaid medical bills & unnecessarily tax public health systems. Solving this issue calls for both legislative actions and a more a dvanced, data-driven early diagnostic and response plan. Predictive analytics has evolved into a powerful tool in public health allowing more organizations to move from reactive to more proactive approaches. By using historical data analysis & pattern identification, it is possible to forecast which members are most prone to lose coverage & to implement quick interventions before such events. While conventional analytics would focus on previous trends, machine learning (ML) would clarify more complex relationships and patterns, therefore most suited for estimating their eligibility gaps. But the most important thing is making sure these technologies are more scalable and easily available given the limits of the public health system. In this sense, cloud-native solutions as Google Big Query ML and Snowflake External Functions are very vital.

With standard SQL searches, Big Query ML lets consumers create & apply ML models straight within Google's BigQuery data database. For data teams without advanced knowledge of Python or another specialty programming language, this lowers the admittance requirements. It guarantees smooth connectivity with healthcare data kept in the cloud & allows the scalable

development of regression, classification, time series & more clustering models. By means of its External Functions feature, Snowflake improves this capability by allowing direct execution of APIs or remote services straight within SQL. Without data crossing platforms, this enables actual time decision-making, the inclusion of third-party eligibility requirements & access to many other data sources including socioeconomic determinants of health. This research intends to show how a predictive pipeline built from an integrated architecture of Big Query ML and Snowflake External Functions may identify their CHIP consumers at risk of coverage termination. Using the simulation of an actual world healthcare dataset, this work shows the training, assessment & implementation of SQL-based ML models to forecast eligibility gaps. Moreover, we show how Snowflake may use outside data to improve their projections and automate flagging processes or outreach depending on risk evaluations. This design underlines the pragmatic advantages of cloud-native, interoperable solutions in public health & shows technical feasibility.

Figure 1. Administrative turnover results from various elements

The work aims to link operational healthcare needs with advanced analytics. Familiar SQL environments allow public health organizations to operationalize their predictive models, therefore removing the need for considerable data wrangling & also software development over many months. The recommended approach underlines the need for adaptability, scalability, and speed qualities more vital for controlling big-scale, dynamic projects like CHIP. Importantly, it shows how public sector initiatives might make use of the same ML techniques transforming companies like marketing, transportation & also finance. Maintaining children's ongoing membership in CHIP serves both a moral & financial need. By using the predictive powers of cloud-native ML, agencies might prevent administrative mistakes, lower turnover & improve outcomes for children depending on their access to care. By providing a practical, repeatable technique that fits the evolving dynamics of healthcare data management and policy innovation, this study promotes those goals.

2. Technical Background & Tools Overview

Technologies enabling huge scale data processing & democratizing access to advanced analytics are desperately needed given the growing volume & more complexity of healthcare data. By using cloud-native, SQL-based ML solutions that allow data teams to create predictive models in familiar environments, this problem may be addressed to forecast eligibility holes in the Children's Health Insurance Program (CHIP). Eliminating the complexity of traditional ML methods, Google Big Query ML and Snowflake External Functions provide complimentary benefits for building scalable, actual time prediction systems. This part provides an overview of these technologies & explains why their combination marks a major development in the analytics field.

2.1 Big Query Machine Learning Based on Native SQL

A strong tool included inside Big Query, Google's corporate data repository, Big Query ML lets users train & implement ML models straight using SQL. By removing the necessity of exporting data to separate ML environments like Python or R, this SQL-native interface helps teams to build, evaluate & apply more predictive models directly within their data warehouse.

2.1.1 Main Features of Big Query ML

Indigenous SQL interface:

To support machine learning activities including CREATE MODEL, ML.TRAIN, ML.EVALUATE, and ML.PREDICT Big Query ML improves regular SQL syntax. Using familiar language, analysts & data engineers can use ML on huge volumes, greatly reducing the learning curve and increasing output.

- Big Query ML fits several model types, hence suitable for many uses.

- Binary classification logistic regression, best for evaluating CHIP eligibility lapses' likelihood.
- For numerical forecasts, linear regression.
- K-means groupings for population segmentation.
- ARIMA_PLus for temporal data prediction.
- Integration of AutoML Tables helps to build more complex models via hyperparameter tuning.
- For latent pattern detection & also suggestions, matrix factorization

Training on huge datasets is more effective as all calculations take place inside Big Query. Crucially for more sensitive healthcare information, this reduces data transmission, lowers latency & maintains their security and compliance. Big Query ML offers methods for evaluating model accuracy (precision, recall, AUC) and analyzing their features using explainable ML capabilities qualities more vital in controlled industries like healthcare.

2.2 Snowflake External Features: Boosting Real-Time Intelligence

Snowflake's External Functions let users contact outside APIs and services straight from inside their SQL systems. Maintaining the continuity of data analysis, this capability is especially helpful for including actual time data or running computations depending on logic outside the data warehouse.

2.2.1 Main Applications for the CHIP Eligibility Project

Integration with External Program Interfaces:

- External purposes could call for services including:
- Procedures of verification for variations in residency or income.
- APIs for instant eligibility confirmation.
- Publicly accessible datasets (such as census data or indexes of Social Determinants of Health).

Feature enrichment and data augmentation help to enhance internal data by means of their outside ideas, hence increasing model correctness. For example, using an API that assesses family stability based on their financial and geographic information can provide fresh, actual time features to improve model predictions. Snowflake may dynamically seek actual time feature extraction when qualifying status might rely on the most current income level or family size, thereby improving the flexibility & currency of the prediction pipeline. Reusability and Scalability: Snowflake's architecture adaptably and promotes great concurrency. Defining External Functions allows one to utilize them across departments, hence promoting consistency in data processing.

2.3 Value of SQL-Based Machine Learning for Public Health

This approach is based on the belief that, not only for individuals with strong programming knowledge but also for all data teams should machine learning be accessible. Data processing across many companies uses SQL as a common language; incorporating ML within SQL environments offers more numerous striking benefits:

2.3.1 Making Machine Learning All Accessible:

- SQL-based ML tools allow data analysts, program managers & domain experts to quickly build and use more predictive models by lowering technical barriers. This addition cuts the time to insight and encourages innovation.
- Data preparation, modeling & deployment may all occur within a single platform in an effortless workflow integration. This strong relationship reduces more operational delays, versioning differences, and handoffs, hence improving their agility in public health analytics.
- For business, data, and IT teams, SQL is a universal language. Combining ML capabilities within SQL helps cross-functional teams to work more effectively in creating therapies for at-risk CHIP groups.
- Sometimes public health choices operate under strict timeframes, accelerating time-to- value. Rapid prototyping & model implementation made possible by SQL-based ML removes infrastructure needs and long development times.

3. Data Sources and Preparation

Great, properly chosen data is the foundation of successful predictive modeling. Forecasting eligibility gaps in the Children's Health Insurance Program (CHIP) requires a sophisticated data engineering technique as the complexity of healthcare data includes enrollment history, demographics & behavioral and socioeconomic characteristics. The basic datasets utilized, the preprocessing techniques used, important feature engineering methods, and the integrated data warehousing pipeline linking Snowflake and Big Query for more effective analytics and ML are described in this section.

3.1 Data Sources

Synthetic simulations closely replicating actual world CHIP enrollment data generated the datasets utilized in this study. These databases mirror the shape and behavior of data routinely maintained by social assistance organizations, health plan managers & state Medicaid and CHIP authorities. Three primary data sources were used: Record of CHIP Enrollment:

- Contains monthly records on child enrollment status.
- Includes re-enrollment events together with timestamps for enrollment's begin and end.
- Monitors arrange categories, renewal initiatives, and, if available, reasons of disenrollment.
- Helps one understand patterns particular to plans, enrollment consistency, and turnover dynamics.

3.1.1 Residential and demographic data:

Covers the age, gender, ethnicity, language preference & more geographic location—ZIP code or census tract—of both the child and guardian. Income categories, family composition, employment status, and access to digital communication tools—such as email or phone are among household-specific data points.

- Helps create personalized risk profiles.
- Administrative Churn Indicators and Event Documentation:
- Points up procedural factors for more coverage loss, including:
- Still waiting for renewal documentation.
- Mail returned because of outdated addresses.
- Not able to turn in documents on time.
- Unsuccessful efforts at electronic verification.
- Finding instances of avoidable turnover and guiding supervised learning models depend on these records.

3.2 Method of Preprocessing

Rarely are raw data ready for modeling. To prepare the data for ML, a sequence of preprocessing tasks was carried out cleansing, alignment & data readiness.

3.2.1 Missing Value Imputation:

- Missing demographic data such as income or ethnicity was imputed using mean or median values at the ZIP code level using many other public databases such as the American Community Survey.
- Absent months were specifically added with a "not enrolled" status to fill in time-series gaps in enrollment history & preserve continuity in longitudinal analysis.
- Administrative flags lacking timestamps were more categorized as non-events rather than nulls to avoid a hyperbole of turnover.

3.2.2 Segmentation and Temporal Coordination:

- Every piece of data was arranged monthly in a panel style, with each row showing the state of a specific kid for a given month.
- Calculated using rolling time frames were lagged variables (e.g., "was enrolled 3 months prior") and retroactive traits.
- Temporal slicing allowed the dataset to be divided into training, validation & test sets based on time—that is, training on data up to 2023 and validating on 2024.

3.2.3 Encoding and Standardization:

- Z-score or min-max normalizing helped to normalize their continuous data like income, family size, and age.
- One-hot encoded, categorical variables included plan type, ethnicity & churn cause codes.

3.3 Feature Extraction

An extensive array of designed features was obtained from the raw datasets in order to improve the prediction power of the models. These included both socioeconomic elements & behavioral inclinations.

3.3.1 Recency, frequency, and lag here Features:

- Recency: The timeliness of a child's enrollment or system interaction that is, the most recent address update.
- Frequency: The number of re-enrollments or turnover events within the last year.
- Lag Variables: Days gone since the last form submission, duration since the most recent eligibility check, etc.

3.3.2 Behavioral References:

- Trends in participation in more administrative processes, including responses to renewal alerts and form submission timings.
- Behavioral proxies generated from digital footprints (e.g., if available, online portal login frequencies).

3.3.3 Socioeconomic Risk Indices:

- ZIP code level data on poverty rates, school food program participation, housing stability & internet accessibility.
- Indices of public health concern historical patterns of disenrollment in underdeveloped economically deprived areas.

3.3.4 Composite Risk Evaluations:

- Created synthetic risk ratings by combining multiple indicators (e.g., restricted digital access, past delayed renewals, and high ZIP-code turnover rate).
- These ratings served as features & a means of classification for risk categories related to possible behavior.

3.4 Pipeline for Data Warehousing: Big Query and Snowflake Integration

To enable the simple data flow & augmentation across platforms, thereby enabling the modeling pipeline, Snowflake and Big Query developed a bidirectional data integration architecture.

3.4.1 Snowflake as the Defining Truth Source:

- Snowflake housed all raw data imports, cleaning & also first transformations.
- Snowflake's scalability and regulation by their governance make it ideal for compiling consistent datasets from various sources.

3.4.2 External Objectives for Improvement:

Snowflake External Functions let actual time searches and feature upgrades run via APIs such as:

- USPS's validation of addresses
- Tools for income validation.
- Public demographic and socioeconomic data sets.
- Big Query for Learning: Machine Learning

Feature-ready tables were either duplicated or federated into Big Query via connectors or scheduled data exports after enrichment. Using the elastic computing of the cloud for massive processing, Big Query ML enabled model training, evaluation & more predictions straight in SQL.

3.4.3 Models and Feedback Mechanism:

Snowflake received prediction outputs and model scores generated in Big Query back-into for:

- Starting alarms or procedures.
- Building reports and dashboards.
- Evaluating actual performance upon application.

4. Modeling Approach

Anticipating eligibility differences in public health efforts including the Children's Health Insurance Program (CHIP) means tackling more numerous complex issues: limited insight into family activities, changing socioeconomic situations & also procedural inconsistencies. Technical accuracy, a rigorous definition of the problem, the choice of suitable models, and the use of cloud-native platforms to apply their insights define a successful modeling strategy. This part clarifies the rationale behind our modeling choices, describes how ML models utilizing Big Query ML are executed & investigates how Snowflake External Functions contribute to increase their prediction accuracy and flexibility.

4.1 Definitions of Problems: Regression versus Classification

Clearly determining the prediction aim marked the first part of the modeling strategy. Two related but distinct problems surfaced within the context of CHIP eligibility:

4.1.1 Binary Classification: Forecasting Churn Risk:

This job includes figuring out if a child enrolled in CHIP today runs risk of losing coverage in the following month or quarter. When the outcome variable is binary 0 (no churn) or 1 churn this is a classic classification issue. Since it lets program managers find & concentrate on high-risk participants for outreach and help, this structure is more suitable for direct intervention

and also policy execution. Time-series forecasting anticipating coverage lapses over time offers and many other perspectives focused on timing & also frequency of enrollment interruptions. Time-series models fit for spotting trends, seasonality & odd patterns in enrollment change over time. Strategic planning & more resource allocation would benefit much from these results. The modeling system offers great help to decision-makers by addressing both tasks: time-series forecasting for strategic planning & more categorization for quick action.

4.2 Model Choosing

4.2.1 Classifier Logistic Regression

For several important reasons, logistic regression became the main model for churn categorization.

- Logistic regression offers a clear interpretation of coefficients, which is necessary for decisions connected to healthcare that call for openness & also justification.
- Scalability: Commonly encountered in administrative systems, it performs quite well on big tabular databases.
- Logistic regression is easy to train, validate & maintain given its few hyperparameters.

In public health environments, models have to be easy for auditing. Because logistic regression is transparent & simple, it complies to regulatory tastes. Our model included engineering features comprising previous churn history, documentation recency, family income level, number of dependents, administrative indicators, and ZIP-code-level risk factors as the independent variables while the dependent variable was churn (yes/no).

4.2.2 ARIMA, Time-Series Prediction,

Overarching enrollment trends & future lapse patterns throughout time were examined using the ARIMA (AutoRegressive Integrated Moving Average) model. Using autoregressive components, differencing to stabilize their trends, and moving averages, this strong statistical method for modeling univariate time series data is presented.

- ARIMA is quite good at estimating overall lapse rates within a given area.
- Understanding seasonal influences that is, annual renewal schedules.
- Planning outreach projects at intervals of critical danger.

We examined aggregated churn data grouped by ZIP code & temporal periods (e.g., month or quarter) using ARIMA models. These results allowed public health officials to forecast rising disenrollment and allocate funds in line.

4.3 Big query ML Execution

Big Query ML helped to develop time-series models & also logistic regression as well. Its ability to run ML operations straight within the SQL-based data warehouse greatly streamlined the process.

4.3.1 Training Model Apply SQL

Training for logistic regression began with a CREATE MODEL command using the specified model type LOGISTIC_REG. From preprocessed & normalized their feature-engineered databases, input features were obtained using a SQL SELECT query.

Training within Big Query offered advantages including:

- Not a data transfer: Every activity happened within the warehouse, guaranteeing security & following policies.
- Elastic computing resources of Google Cloud were applied sensibly without human provisioning.
- Efficiency and Ease: SQL-savvy analysts may quickly begin & polish models without Python's or R's coding required.

Time-series predictions were trained using the ARIMA_PLUS model. Big Query ML determined appropriate hyperparameters depending on the training set after independently spotting seasonality.

4.3.2 Model Validation and Optimization:

Validation achieved using:

- Retention Term Windows, Microsoft: Temporal criteria that is, data before 2020 for training, then thereafter data for testing separated the dataset into training & testing sets.
- For classification, we evaluated precision, recall, F1-score, and AUC. ARIMA model accuracy was assessed using RMSE (root mean square error) & MAPE (mean absolute percentage error).
- Though logistic regression contains less hyperparameters than more complex models, tests were run involving feature selection, regularization techniques & also more interaction terms.

4.4 Expanding with External Snowflake Functions

Snowflake External Functions were added into the pipeline for workloads beyond SQL's intrinsic capability in order to further the predictive architecture.

- Python User-Defined Functions for more Complex Feature Extraction
- External functions let Python-based User-Defined Functions (UDFs) be invoked within Snowflake SQL scripts to extract more complex traits from unstructured or semi-structured data, including text-based comments from contact center logs.
- Calculate entropy scores to evaluate behavioral variability.
- Add outside services for community health index values, economic measurements, or ZIP-code risk assessments.

The enhanced features were then combined with the main dataset and made available for Big Query ML, therefore improving the feature space and increasing the model performance.

4.4.1 Advanced Modeling Made Possible by Survival Analysis

Often employed in more clinical research, survival analysis was used to replicate the length until turnover takes place. Big Query ML does not natively support it, however outside utilizing Snowflake External Functions a Python implementation was done. This method provided some very insightful analysis on median enrollment duration.

- Hazard ratios throughout different populations.
- Churn probability over many timescales.

These outputs were particularly helpful for organizing intervention efforts in line with temporal urgency.

4.4.2 Snowflake Reenter Result Integration

After Big Query ML generated model predictions, Snowflake received the results—risk scores, churn probability & time-series forecasts via either scheduled data exports or data replication techniques.

- These outputs in Snowflake were used to begin automated procedures (such as flag notes for caseworker review, alert distribution).
- Displayed on dashboards for CHIP program managers.

Combined with outside factors (such as seasonal outreach plans, financial cycles) for coherent decision-making.

5. Case Study: Predicting Coverage Gaps in a State CHIP Program

This case study investigates and corrects enrollment discrepancies by means of a simulated predictive modeling implementation into a state-level Children's Health Insurance Program (CHIP). By closely matching the structure, variables & more operational logic of actual CHIP datasets, the synthetic data employed helps us to show a useful application free from more reliance on private or sensitive information. This case study seeks to define the methodical building of a ML pipeline to prioritize their vulnerable children, evaluate the effectiveness of proactive outreach operations & forecast short-term coverage failures.

5.1 Overview of Simulated State-Level CHIP Information

The data utilized in this simulation relates to a mid-sized U.S. state's CHIP program, with around 150,000 children spread over a five-year period (2016–2020). Together with demographic, socioeconomic & more administrative interaction information, it comprises monthly records of enrollment, disenrollment & also re-enrollment.

5.1.1 The collection consists of necessary tables including:

- Monitors each child's monthly enrollment status including gaps, renewals & related timestamps in an enrollment history table.
- Household Data Table: Shows household size, digital access (email or phone), zip code, guardian income group, and language of choice.
- Records renewals, failed verifications, missed documentation deadlines & churn cause codes in an administrative log table.
- Enhanced with publicly accessible ZIP code-level data including poverty rates, housing instability & educational indices, Geographic Risk Indicators

This dataset was generated to evaluate whether a ML model could correctly estimate which children will lose coverage over a three-month period & if such forecasts may direct focused actions.

5.2 Methodical Analysis

5.2.1 Feature Set Specification

The initial step consisted in the creation of a complete feature set produced from the current data. Five categories comprised the features:

5.2.1.1 Patterns of enrollment:

- Count of shortages throughout the last twelve months.
- Timeliness of most recent re-enrollment.
- Length of continuous enrollment term.

5.2.1.2 Administrative Involvement:

- Timeliness of obligatory document submission.
- Total count of failed correspondence.
- Indices show defunct phone lines or undeliverable mail.

5.2.1.3 Residential and Demographic Context:

- The child's age and gender.
- Dimensions of the house and income category.
- Guardian's employment status.

5.2.1.4 Geospatial and Socioeconomic Vulnerability:

- Zip code level health disparities index.
- Accessibility to broadband internet.
- Closeness to public help centers.

5.2.1.5 Composite Risk Indicators:

- Consolidated behavioral risk score drawn from past events.
- Flag for seasonal renewal planning, say for academic enrollment or vacation time.

Monthly updates for all traits created a longitudinal panel allowing the model to learn over time from more behavioral patterns and also temporal dynamics.

5.2.2 Modelling Training

Big Query ML was used to build a logistic regression model to predict if a child will have a coverage gap over the next ninety days (binary target variable: 1 = lapse, 0 = no lapse).

- The training period is 2016–2019.
- Period of Validation/Test: 2020.
- Data Granularity: About 9 million records one record per child per month.

The logistic regression classifier was built straight in SQL by using Big Query ML's CREATE MODEL and ML.TRAIN commands. Big Query's SQL capabilities allowed all preprocessing tasks including normalizing, categorical variable encoding & more temporal lag feature generation executed upstream. Grid search and 5-fold cross-valuation on the training set helped to tune hyperparameters of the model that is, regularization strength.

5.2.3 Confusion Matrix and ROC-AUC

The model's generalization capacity to fresh events was evaluated with 2020 holdout data after training. Many necessary steps were taken:

- Area under curve - receiver operating characteristic
- With an area under the curve (AUC) of 0.84 the model showed strong discriminative power between those who would and would not suffer coverage loss.

5.2.3.1 Precision and sensitivity:

- Precision: 78% the proportion of indicated children who really had a lapse.
- Recall: 71% That is, the proportion of actual mistakes precisely noted.

- The confusion matrix showed a reasonable faulty positive rate, judged reasonable given the great cost involved in a wasted outreach opportunity.

Graphs of probability calibration showed that the anticipated probabilities of the model matched the observed lapse rates exactly. The results showed that in public health settings, high-quality, painstakingly gathered data may provide valuable insights even using a basic model like logistic regression.

5.3 Applications

5.3.1 Projecting Lapse Probability for the Next Three Months

The model ran on the active enrollment group monthly to give every child a three-month lapse probability score. Children with a predicted risk higher than a certain threshold say, 70% were found for assessment & more possibly intervention. Via scheduled queries from Big Query, the forecasting & more prediction tasks were automated & connected with downstream systems via Snowflake, therefore enabling administrative staff to see model results in actual time.

5.3.2 Determining Outreach Risk Children

The highest quartile of expected at-risk patients was channeled into a process activating intervention plans involving email or reminder message distribution.

- Assigning direct phone communication caseworkers.
- Sending renewal materials ahead of time along with being simplified by their guidelines.
- Offering household profiles specific language assistance.

Every outreach project was recorded to evaluate more response rate and consequent impact on coverage continuity.

5.4 Outcomes and Influences

After simulating a year-long intervention period based on 2020 test data, we evaluated the model's actual world impact retrospectively.

5.4.1 Reducing Gap Rates

- Pre-model initial gap rate: 14.3%
- Ten-1% simulated post-intervention gap rate
- Relative reduction: around 29%

Using a risk-based outreach approach helped to drastically reduce the general turnover rate. It also improved coverage continuity for children from historically underprivileged ZIP codes and among homes without internet access.

5.4.2 Economic Analysis

We computed the running expenditures of outreach in relation to the savings obtained by preventing lapses to assess economic viability:

- Projected child outreach expense: $3.50
- Coverage Lapse Expense: About $300 each case, including ER visits and reprocessing
- For every dollar allocated to outreach, roughly $6.75 was saved for administrative and healthcare expenses.

Furthermore supporting the expenditure in predictive modeling and outreach programs were intangible benefits like improved health outcomes and decreased stress for families.

6. Conclusion

This work presents a practical & more effective method using scalable, SQL-driven ML tools for estimating coverage deficits in the Children's Health Insurance Program (CHIP). To enable the ingestion, processing, modeling & more deployment of predictions totally within structured data environments, we created a cloud-native solution combining Google Big Query ML with Snowflake External Functions. Our approach included pretreatment of enrollment and demographic data, engineering of important components like churn history and socioeconomic risk, and use of logistic regression for individual churn prediction with ARIMA models for time-series lapse forecasting. The case study showed that a logistic regression model built in Big Query ML had a strong ROC-AUC of 0.84, displaying high accuracy and recall in identifying children at risk of disenrollment. These projections encouraged targeted participation that, in simulation, reduced gap rates by around 29% & produced significant price savings. By

means of Snowflake additional functions, actual time feature improvement and interaction with many other data sources including geographic indexes and verification APIs was enabled, thereby boosting model accuracy & more practical relevance.

The results provide strong evidence of the possibilities SQL-based ML offers for public sector analytics. While conventional ML pipelines may call for expert teams and diverse technology, Big Query ML lets analysts create models within their present processes using known SQL syntax. This assures that important insights stay close to the data, speeds development, and reduces many barriers to access. Snowflake benefits the same advantages wherein external function calls provide dynamic, real-time decision-making while preserving governance and scalability. Apart from its technical advantages, this integrated analytics system has strategic relevance for public projects. Early treatments, best use of resources & evidence-based policy decisions all vital for improving outcomes in programs like CHIP are made easier by it. By being able to forecast danger in almost actual time, authorities may respond before an event occurs rather than just after it. The importance of democratizing ML in government analytics is underlined by this paper. We can ensure that at-risk groups especially children retain ongoing access to essential treatment & help better informed public health decisions by giving analysts & program administrators user-friendly tools & intelligible models. This marks not just a technological development but also a step toward a more equal and responsive public service structure.

References

[1] Dageville, Benoit, et al. "The snowflake elastic data warehouse." *Proceedings of the 2016 International Conference on Management of Data*. 2016.

[2] Mucchetti, Mark. *BigQuery for Data Warehousing*. 2020.

[3] Ali Asghar Mehdi Syed. “Impact of DevOps Automation on IT Infrastructure Management: Evaluating the Role of Ansible in Modern DevOps Pipelines”. *JOURNAL OF RECENT TRENDS IN COMPUTER SCIENCE AND ENGINEERING (JRTCSE)*, vol. 9, no. 1, May 2021, pp. 56–

[4] Goss, Raymond, and Lokesh Subramany. "Journey to a Big Data Analysis Platform: Are we there yet?." *2021 32nd Annual SEMI Advanced Semiconductor Manufacturing Conference (ASMC)*. IEEE, 2021.

[5] Jain, Shrainik. *Learning from SQL: Database Agnostic Workload Management*. Diss. 2019.

[6] Cloud, Securing Your Snowflake Data, Ben Herzberg, and Yoav Cohen. "Snowflake Security."

[7] Atluri, Anusha. “The Autonomous HR Department: Oracle HCM’s Cutting-Edge Automation Capabilities”. *International Journal of Emerging Trends in Computer Science and Information Technology*, vol. 3, no. 1, Mar. 2022, pp. 47-54

[8] Ronthal, A. M., Roxane Edjlali, and Rick Greenwald. "Magic Quadrant for Data Management Solutions for Analytics." *Gartner, Inc. ID: G00326691* (2018): 1-39.

[9] Anand, Sangeeta, and Sumeet Sharma. “Hybrid Cloud Approaches for Large-Scale Medicaid Data Engineering Using AWS and Hadoop”. *International Journal of Emerging Trends in Computer Science and Information Technology*, vol. 3, no. 1, Mar. 2022, pp. 20-28

[10] Tang, Chunxu, et al. "Forecasting SQL query cost at Twitter." *2021 IEEE International Conference on Cloud Engineering (IC2E)*. IEEE, 2021.

[11] Beygenov, Askhat, et al. "Audubon Data Project Final Report." (2018).

[12] Armbrust, Michael, et al. "Delta lake: high-performance ACID table storage over cloud object stores." *Proceedings of the VLDB Endowment* 13.12 (2020): 3411-3424.

[13] Ali Asghar Mehdi Syed, and Shujat Ali. “Evolution of Backup and Disaster Recovery Solutions in Cloud Computing: Trends, Challenges, and Future Directions”. *JOURNAL OF RECENT TRENDS IN COMPUTER SCIENCE AND ENGINEERING (JRTCSE)*, vol. 9, no. 2, Sept. 2021, pp. 56-71

[14] Miranda, Serge. "FROM DATA BASE TO BIG DATA MANAGEMENT." (2019).

[15] Kenney, Genevieve M., et al. "Children eligible for Medicaid or CHIP: who remains uninsured, and why?." *Academic Pediatrics* 15.3 (2015): S36-S43.

[16] Vasanta Kumar Tarra, and Arun Kumar Mittapelly. “Future of AI & Blockchain in Insurance CRM”. *JOURNAL OF RECENT TRENDS IN COMPUTER SCIENCE AND ENGINEERING (JRTCSE)*, vol. 10, no. 1, Mar. 2022, pp. 60-77

[17] Gresenz, Carole Roan, et al. "Income eligibility thresholds, premium contributions, and children's coverage outcomes: a study of CHIP expansions." *Health Services Research* 48.2pt2 (2013): 884-904.

[18] Atluri, Anusha. “Extending Oracle HCM Cloud With Visual Builder Studio: A Guide for Technical Consultants ”. *Newark Journal of Human-Centric AI and Robotics Interaction*, vol. 2, Feb. 2022, pp. 263-81

[19] Saloner, Brendan, Stephanie Hochhalter, and Lindsay Sabik. "Medicaid and CHIP premiums and access to care: a systematic review." *Pediatrics* 137.3 (2016).

[20] Yasodhara Varma Rangineeni. “End-to-End MLOps: Automating Model Training, Deployment, and Monitoring”. *JOURNAL OF RECENT TRENDS IN COMPUTER SCIENCE AND ENGINEERING (JRTCSE)*, vol. 7, no. 2, Sept. 2019, pp. 60-76

[21] Brooks, Tricia, et al. "Medicaid and chip eligibility, enrollment, renewal, and cost sharing policies as of january 2017: Findings from a 50-state survey." *Kaiser Family Foundation Report* (2017).

[22] Pei, Zhuan. "Eligibility recertification and dynamic opt-in incentives in income-tested social programs: Evidence from Medicaid/CHIP." *American Economic Journal: Economic Policy* 9.1 (2017): 241-276.