**Ebill Enrollment Model for Residential Customers**

By Alec Zhixiao Lin

September 2025

**Contents**

1. Purpose of the model
2. Data Sources
3. Experimental Design
4. Data Extraction
   1. Modeling population
   2. Aggregating behavioral data
   3. Combining data for modeling
      1. Modeling sample
      2. Validation sample
5. Variable Evaluation and Selection
6. Modeling
7. Model Deployment
8. Model Performance Monitoring
9. Some Afterthoughts
10. **Purpose of the model**

Currently around 64% of residential customers have been enrolled in paperless billing. SCE intends to increase this enrollment rate to 67% within a year.



**Table 1 – Summary of bill types for residential customers**

The summary above and the model to be built are at the level of ContractAccount. Customers receiving Hard Copy Bill or with missing Bill Type (both in yellow highlights) will be selected as the population for modeling if they meet the selection criteria (see 4.1).

**2. Data sources**

Data is extracted from the following three sources:

1. Snowflake: PROD\_CS\_SS.CS\_CSOD\_BIC\_SS.CS\_EVENTS

This table stores transactional events for all customers. It includes dates and associated amount in dollars (if any).

2) SAP Hana: multiple tables are used to extract data from. See python programs in 4.1 for all tables used.

3) Acxiom: socio-demographic information for residential customers.

**3. Experimental design**

We used customers’ profile and behavioral data from the past year as predictor variables (X). E-bill enrollment within 30 days following the cut-off date serves as the target variable (y) for modeling.

For building modeling sample and validation sample, we applied the following experimental design:



**Table 2 – Experimental Design**

The predictors include profile information such as account age, rate plan, etc. Behavioral data will be aggregated to the level of ContractAccount for modeling.

We need to extract or reconstruct a historical snapshot that reflects active customers on a specific day. However, reconstruction is extremely time-consuming or nearly impossible. Fortunately, on June 26, 2025, we executed the SAS code that generated a file containing active customers as of that day (see 4.1). On the day of August 11, 2025, we executed the same SAS code that generated a similar file containing all active customers on that day. This file is to be used to build a validation sample.

**4. Data Extraction**

The following sections outline the steps taken to construct the modeling sample. These same steps apply to building a validation sample, with the only difference being the cut-off date.

**4.1 Modeling population**

*Program used: step1-modeling sample 20250626 cleaned.ipynb*

At the start of building the model, we need to find a snapshot of all residential customers who meet the following criteria:

* Active on the day of data pull.
* Not enrolled in E-bill
* Eligible for enrollment[[1]](#footnote-1)

The following SAS program was executed on June 26, 2025 pulled customers active that day:

A screen shot of a computer code

AI-generated content may be incorrect.

Today() in the above SAS code refers to 2025-06-25 as it was the date when this program was run. The output SAS file was uploaded to GCP as *cact\_not\_in\_ebill\_0626.csv* for merging.

The rest of the program mainly uses python for data extraction and processing.

* 1. All residential customers active as of 06/26/2025
  2. All customers with emails as of 06/26/2026

Emails were pulled from two sources[[2]](#footnote-2):

CV\_CONSENT\_GUID\_WEBID,

SV\_BP\_CONTACT\_DETAILS

The data extracted above provides profile information at the ContractAccount level. Later, we will derive Account Age as a predictor by calculating the difference between the customer's Move-In Date and the cut-off date, which in this case is June 26, 2026.

**4.2 Aggregating behavioral data**

*Program used: step1-modeling sample 20250626 cleaned.ipynb*

(The section numbers below correspond to section numbers in the markdowns in the python program.)

Behavioral data was extracted from both SAP HANA and Snowflake. The majority of this data consists of customers’ behavioral information from the past year[[3]](#footnote-3). Since these tables retain event types and associated timestamps, we need to aggregate the data to the customer level for modeling purposes.

Due to the large volume of certain datasets, some data had to be extracted in partitions.

The following steps outline data to be extracted from SAP Hana:

* 1. – customers who have enrolled and opted later before the cut-off date.
  2. – Payment information in last year[[4]](#footnote-4)

Due to its large volume, we extracted in partitions by 30 days, 60 days, 90 days, 180 days and 365 days. Payment recency is also extracted.

* 1. – Enrollment and opt-out after cut-off date

Data pulled in this step will be used as target variable y for modeling. Those customers who opted out after enrollment will not be included in y.

The Transactional Insights Platform (TIP) is the sole source from which behavioral data was extracted in Snowflake. The following steps outline the process used to extract data from Snowflake:

2.1.1 -  ContractAccount Level count variables for event type

2.1.2 - ContractAccount Level recency for event type

2.1.3 - ContractAccount Level amount (in dollar) variables for event type

2.2.1 - ContractAccount Level count variables for subtype

2.2.2 – recency for subtype

2.2.3 – Amount for subtype

2.3.1 – count variables for event+subtype

2.3.2 – recency variables for event+subtype

2.3.3 – Amount variables for event+subtype

All extracted files will be stored in Google Cloud Storage to be merged later.

**4.3 Combining all data for modeling**

Since the modeling and validation samples are based on different cut-off dates, they were extracted separately. The following sections detail the steps used to pull data for the modeling sample. The same steps apply to the validation sample, with the only difference being the cut-off date.

**4.3.1 Modeling sample**

*Program used: step1-modeling sample 20250626 cleaned.ipynb*

This refers to Part 3 of the above python program. It does the following merging sequentially:

* Merge all residential customers with their email addresses. Only customers with emails are eligible for Ebill enrollment.
* Merge with new enrollment (6/26/2026-7/25/2025)
* Merge with payment data extracted from SAP Hana
* Merge with behavioral data extracted from Snowflake.
* Merge with Acxiom data

After reading in Acxiom data, we need to modify the variables by removing or changing special symbols contained in the column names.

At the end of the program a file names *model\_sample2.csv* is constructed as the modeling sample with around 1000 potential predictors.

**4.3.2 Validation sample**

*Program used: step5.1-validation sample data pull 20250811 cleaned.ipynb*

The above program contains exactly the same steps as in *step1-modeling sample 20250626 cleaned.ipynb* for building the modeling sample, except that the cut-off date is changed to 08/11/2025.

**5. Variable evaluation and selection**

*Program used: var\_redu\_ebill2.sas, b1.appendix ra.sas, b2.appendix rb.sas*

The modeling sample is imported into SAS for variable evaluation. Information Value (IV) is used to rank the predictive power of each attribute. Variables with higher IV scores—indicating stronger predictive power—are selected for the next step of correlation analysis. This step is run only once. Once variables have been evaluated and selected, we do not need to rerun this.

For output files, refer to *iv\_mod\_y*, which lists all variables ranked from highest to lowest by their IVs. You can also review *gh\_mod\_y.pdf* to examine the behavior of each variable. In the following two examples, the variable on the left suggests a higher predictive power than the one on the right:

 

For theoretical foundation of the methodology, see my paper [095-2013: Variable Reduction in SAS® by Using Weight of Evidence and Information Value](https://support.sas.com/resources/papers/proceedings13/095-2013.pdf) published in SAS Global Forum 2013.

**6. Modeling**

*Program used:* *step5.2 - validation 20250811 cleaned.ipynb*

This program does the following:

* Among 1000 predictors evaluated, selected top 200 with higher predictive power
* Impute for missing values
* Correlation Analysis. Variables with high correlations with other variables of higher predictive power are to be removed.

All machine learning methods for classification have been tried. They include the following:

* Logistic regression
* Decision Tree
* Naïve Bayes
* Stochastic gradient descent
* KNN
* Light gradient boosting machine (LGBM)
* XGBoost
* Random Forest

Since LGBM and XGBoost demonstrated the strongest performance based on AUC-ROC scores, these two models were selected for further analysis[[5]](#footnote-5). These two scores have a correlation of 0.82. After conducting multiple experiments with various combinations of their scores, the following composite score yielded the best monotonicity in terms of predictive performance.:

Propensity\_score=(probability\_LGBM x probability\_XGBoost)1/2

The above formula retains the probability-like estimates.

The following is the model performance:

A graph with numbers and a line

AI-generated content may be incorrect.

**Figure 1 – Model performance**

**7. Model Deployment**

*Program used: step6-implementation data pull cleaned.ipynb*

This program follows the same steps as the above two for generating the modeling sample and the validation sample, with the cut-off date defined as the current date.

For scoring, the scoring file is labeled as seg=’val’ so that the algorithms of LGBM and XGBoost can be applied by scikitlearn of python.

In the Python-based model deployment workflow, we plan to transition the data extraction for active customers from SAP HANA to Snowflake. Snowflake offers the advantage of being directly linkable to the email table. By joining these two tables, we can retrieve all active customers with associated email addresses in a single query, significantly reducing processing time.

**8. Model performance monitoring**

*Program used: step6-implementation data pull cleaned.ipynb*

This part is still being worked on. The following is the preliminary results from examining the enrollment based on the deliverable list generated on September 9, 2025. Please note that we only have 1-2 days of performance data for this list.

A graph with a line going up

AI-generated content may be incorrect.

**Figure 2 – Model performance**

1. **Some afterthoughts**

Acxiom data has provided strong lift for the model. However, SCE’s license with Acxiom is set to expire in December 2025. Starting January 2026, we will transition to data from a new vendor, and the model will be refreshed accordingly.

Customers’ behavioral information has proven to be highly predictive. With the continued expansion of the Transactional Insights Platform (TIP) in Snowflake, we will be able to leverage even more behavioral data to further enhance model performance.

The same aggregated behavioral data can be leveraged to develop a comparable enrollment propensity model for non-residential customers. Deployment can be executed in parallel with the residential customer model. (This effort is already underway.)

1. Business team typically applies several exclusion flags - such as senior citizens, employee accounts, and customers with recent default records - to identify customers eligible for enrollment. However, for this model, they recommend not applying these exclusions to maintain full flexibility in segmentation for a later stage. Email availability will continue to be a key criterion. [↑](#footnote-ref-1)
2. In September 2025, I was pointed to PROD\_CS\_SS/CS\_CSOD\_BIC\_SS/OC\_BP\_EMAIL in Snowflake as a better source for emails. I ran an analysis and found that email-related tables Snowflake and SAP Hana have an overlap close to 100%, with the Snowflake table PROD\_CS\_SS/CS\_CSOD\_BIC\_SS/OC\_BP\_EMAIL showing 400 more emails. Therefore, model deployment will use PROD\_CS\_SS/CS\_CSOD\_BIC\_SS/OC\_BP\_EMAIL for the ease of data extraction. [↑](#footnote-ref-2)
3. Due to compliance considerations, we do not use information that is more than one year old. Additionally, data in both SAP Hana and Snowflake dates back to April, 2021, when marks the timing of CSRP migration. [↑](#footnote-ref-3)
4. There are multiple views in SAP Hana that record customers’ past behaviors. However, a majority of these tables lack clear definition and very time-consuming to extract. Payment table is the one that we can work with. Due to the very long time needed for extraction, we do not include amount-related information in the data extraction. [↑](#footnote-ref-4)
5. The decision to enroll in E-bill may be influenced by numerous factors not captured by the selected model attributes. Additionally, with an enrollment rate of only 3–4% among residential customers, it represents a relatively rare event. As a result, common classification metrics such as accuracy, precision, recall, and F1 score may not be the most appropriate for evaluating model performance. Instead, metrics like lift charts or gains charts are better suited for assessing effectiveness in this context. [↑](#footnote-ref-5)