

Salary Earner Identification and Prediction

1. Introduction

This report details the process documented in the salary.html notebook for identifying customers receiving regular monthly income (likely salary) and predicting their future monthly salary amount using transaction data. The methodology combines rule-based heuristics derived from financial domain knowledge with machine learning techniques for prediction.

2. Data Source

The analysis utilizes transaction data sourced from the `customer_account_transaction_hx` table within a PostgreSQL database. This table contains essential transaction details like `accountid`, timestamps (`trx_start_date`, `trx_end_date`), `amount`, `trx_type`, `trx_subtype`, `initiated_by` (renamed from `d3`), and `description`.

3. Methodology - Identifying Salary Earners

A multi-hypothesis approach was employed to identify potential salary earners by analyzing credit ('C') transactions.

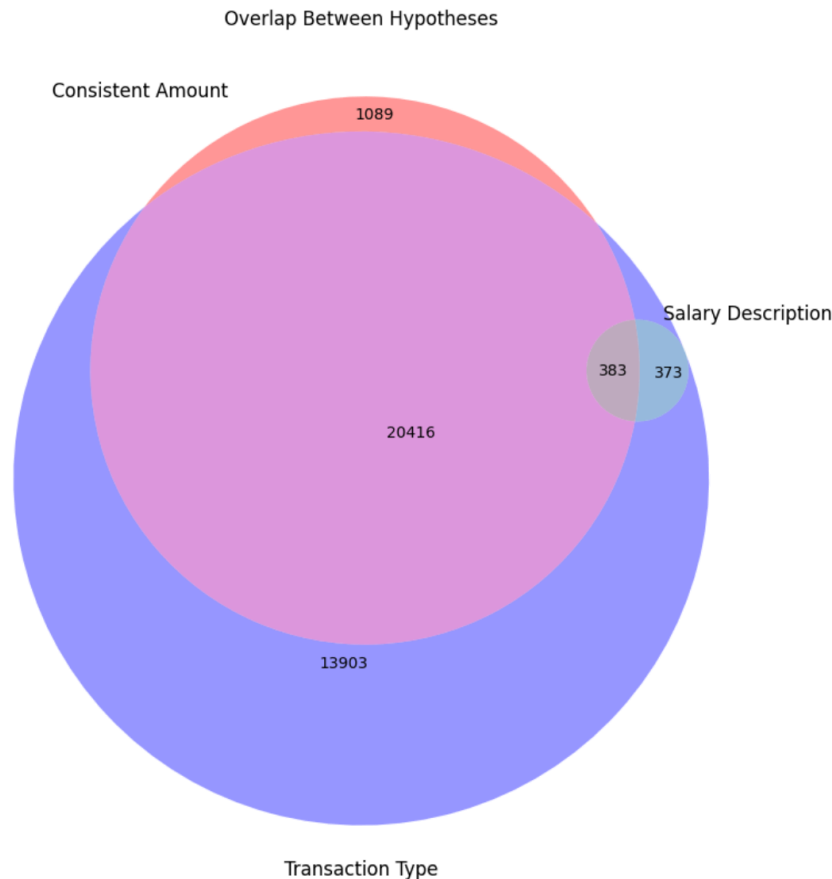
Hypotheses Used:

- I. **Keyword-Based (Hypothesis 1):** Transactions where the `description` field contained specific keywords (e.g., "salary", "payroll", "allowance", month names like "MARCH") and `initiated_by` was 'C'.
- II. **Regular Monthly Intervals (Hypothesis 2):** Transactions occurring at roughly monthly intervals (median interval +/- 15% tolerance) for a given account, focusing on 'C' transactions.
- III. **Consistent Transaction Amounts (Hypothesis 3):** Accounts exhibiting low variability (Coefficient of Variation ≤ 0.10) in the `amount` of their 'C' transactions.
- IV. **Transaction Type/Subtype/Initiator Combination (Hypothesis 4):** Transactions matching specific `trx_type` ('T' or 'C'), `trx_subtype` ('BI', 'I', 'BS', 'CI'), and `initiated_by` ('C'), with a positive amount.

The notebook analyzed the overlap between **Hypothesis 1 (Keywords)**, **Hypothesis 3 (Consistent Amounts)**, and **Hypothesis 4 (Transaction Type/Subtype)** using a Venn diagram. *Note: Hypothesis 2 (Regular Intervals) was defined but appears excluded from this specific final overlap analysis shown.*

Verified Salary Earners: Accounts flagged by *all three* of these hypotheses (Keyword, Consistent Amount, Transaction Type) were classified as verified. The notebook found **383** such unique accounts.

Likely Salary Earners: Accounts flagged by *two out of the three* hypotheses used in the Venn diagram were classified as likely salary earners. The notebook identified **20,797** such unique accounts (this group also includes the verified earners).



A **Venn Diagram** visually represented the overlap between the accounts identified by Hypotheses 1, 3, and 4.

4. Salary Earner Details Table

For both "Verified" and "Likely" salary earner groups, summary tables were generated containing the following key information per `accountid`:

- `accountid`: The unique account identifier.
- `num_months`: The number of months for which salary/income-like transactions were observed for that account within the analyzed data subset.
- `least_inflow_6m`: The minimum credit transaction amount observed for the account in the last 6 months of the data.
- `avg_monthly_salary`: The average amount of the identified salary/income-like credit transactions for that account.
- `estimated_next_amount`: The predicted salary amount for the next expected payment (calculated as the `avg_monthly_salary`).
- `estimated_next_date`: The predicted date for the next salary payment (calculated by adding the median interval between past payments to the last observed payment date).
- `45daysalary` (Boolean): `True` if the last observed salary payment for the account was within the last 45 days from the time of analysis (relative to the notebook's execution time or data snapshot date).
- `2monthssalary` (Boolean): `True` if the account had at least two salary payments recorded within the last 60 days of the data.

5. High Earner Analysis ($\geq 10k$)

An analysis was performed specifically on the **Verified Salary Earners** to identify those with a high average monthly salary.

Criterion: avg_monthly_salary >= 10,000.

Result: 307 verified salary earners met this criterion. A table (high_earner_details_df) was generated showing the accountid and the least_inflow_6m for these high earners.

6. Methodology - Salary Prediction using Machine Learning

To predict the next monthly salary inflow amount, Machine Learning models were developed separately for the "Verified" and "Likely" salary earner groups.

Objective: Predict the average monthly salary amount (amount) for the subsequent month based on historical patterns

Model Choice: RandomForestRegressor from scikit-learn was used.

Feature Engineering:

- Extracted month from trx_start_date.
- Created month_seq (a sequential month counter per account).
- One-hot encoded the categorical trx_type feature.
- Calculated 3-month rolling sum (rolling_sum_3m) and average (rolling_avg_3m) of transaction amounts per account.

Data Preparation & Splitting:

Data was filtered for the relevant account group (Verified or Likely). Transactions were aggregated monthly using the engineered features. Accounts with fewer than 12 months of data were excluded. A time-based split using a sliding window approach was implemented:

- **Training:** Data from the first 8 months (specifically, sequences ending at months 5, 6, 7, and 8, using the preceding 5 months as input features) was used for training.
- **Testing:** Data from the subsequent 4 months (sequences ending at months 9, 10, 11, and 12, using the preceding 5 months as input features) was used for validation.

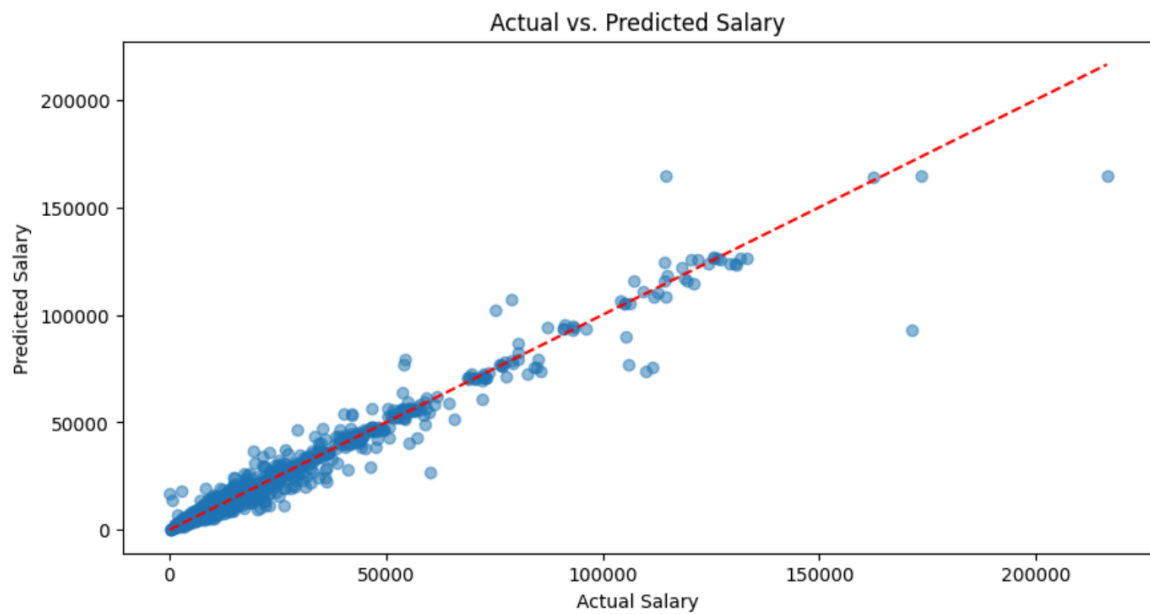
StandardScaler was applied to the feature sets (X_train, X_test) before model training and prediction.

7. ML Model Results

Models were trained and evaluated for both groups:

1. Consistent (Verified) Salary Earners:

- **Accounts Modeled:** 376 (after filtering for >= 12 months data).
- **Performance:**
 - MAE: 1909.79
 - RMSE: 4610.21
 - R-squared: 0.96



2. Likely Salary Earners:

- **Accounts Modeled:** 20,614 (after filtering for ≥ 12 months data).
- **Performance:**
 - MAE: 3209.99
 - RMSE: 51406.22
 - R-squared: 0.97

