CreditKarma Suite

Jennifer Poernomo

MOTIVATION

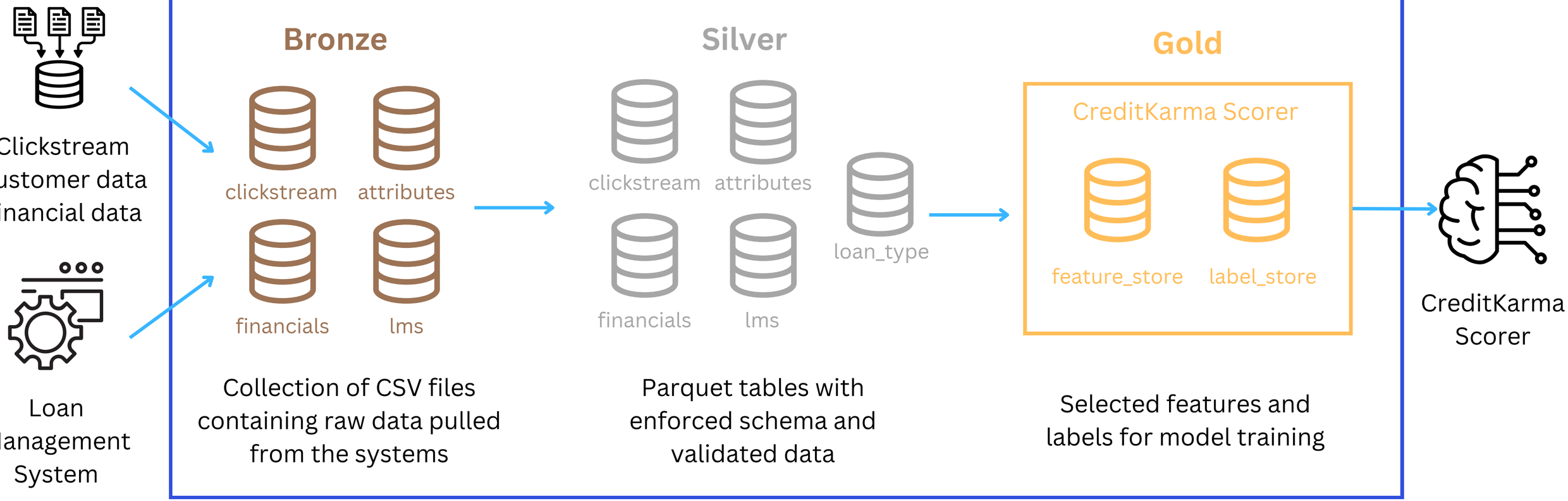
Utilize historical data to enhance loan decision-making and minimize the risk of financial loss from high-risk customers.

CREDITKARMA SUITE

A unified ecosystem of tools that aggregates data from the organization's diverse systems to assess and analyze customer credit scores.

CREDITKARMA DATA LAKE

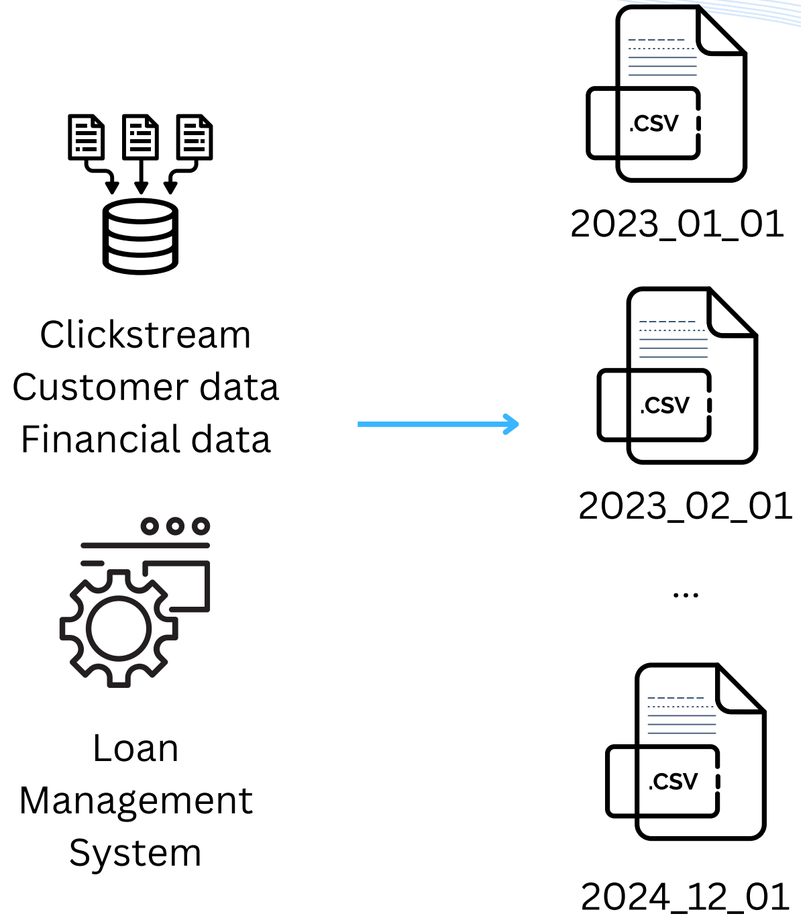
Follows the Medallion Architecture for consistency across applications and ease of maintainability



BRONZE TIER

PARTITIONED CSVS

- Functions as a single source of truth that future CreditKarma maintainers can refer to in case new silver tables need to be built
- Consolidates raw data stream from systems into CSV files partitioned by **snapshot_date**
  - Partitions allows us to query faster & easily retrieve recent data



SILVER TIER

ENFORCED SCHEMAS

- Enforce data types of columns for each table
- Standardized columns to lower case
- Stores each table as a PySpark parquet file
  - Enhances speed of querying and reduces file size due to internal compression

Additional Data Processing

- **Financials:**
  - Split Credit_History_Age (originally string data type) into credit_history_age_year (int) and credit_history_age_month (int)
  - Split Type_of_Loan into a separate table (loan_type) consisting of frequency columns: mortgage_loan, auto_loan, credit-builder_loan, personal_loan, not_specified, student_loan, home_equity_loan, payday_loan, debt_consolidation_loan

SILVER TIER

REMOVED INVALID VALUES

attributes

	Age
count	530.000000
mean	115.375472
min	-500.000000
25%	25.000000
50%	34.000000
75%	43.000000
max	8547.000000
std	775.440211

→ **age < 0 → NULL**

Rationale: Might have children/students who open joint accounts, cannot infer business policy from data alone

→ **age > 120 → NULL**

Rationale: Oldest recorded person is around 120 years old.

Imputation is not done at this stage as the CreditKarma Dashboard might want to show null data & future models might require different methods of imputation.

Enforce SSN to follow AAA-GG-SSSS, NULL otherwise

	Customer_ID	Name	Age	SSN
21	CUS_0x16f4	Forgionez	37	#F%$D@*&8
41	CUS_0x1c9c	Dougq	28	#F%$D@*&8
50	CUS_0x2297	Sergio Goncalvesc	43	#F%$D@*&8
74	CUS_0x2cae	Dunaiu	28	#F%$D@*&8
94	CUS_0x34fa	Carolined	35	#F%$D@*&8

Treat empty as NULL

```
array(['Accountant', 'Developer', 'Lawyer', 'Manager', 'Doctor',
       'Mechanic', 'Journalist', 'Media_Manager', 'Teacher',
       'Entrepreneur', 'Writer', 'Musician', 'Engineer', '_______',
       'Scientist', 'Architect'], dtype=object)
```

SILVER TIER

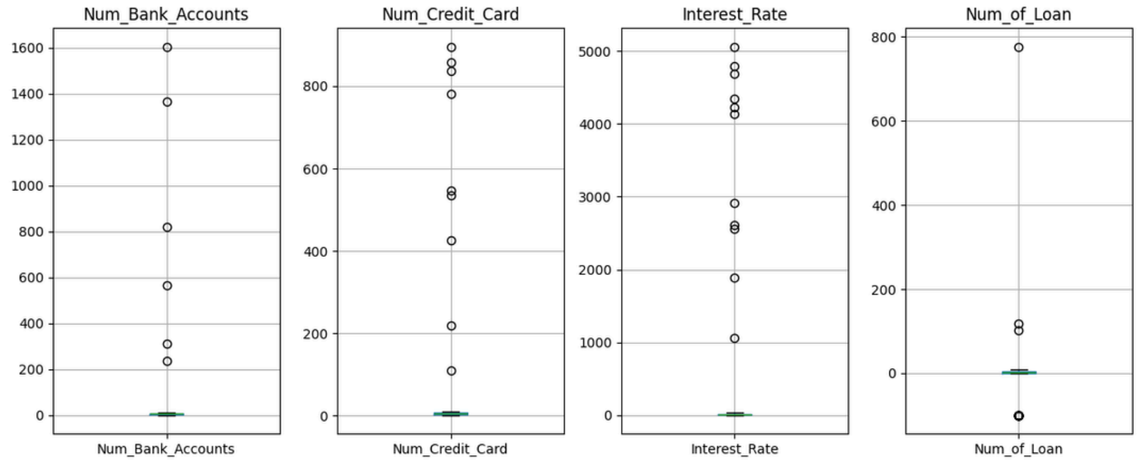
REMOVED INVALID VALUES

financials

	Num_of_Loan	Delay_from_due_date	Num_of_Delayed_Payment
count	530.000000	530.000000	530.000000
mean	0.250943	20.716981	30.464151
min	-100.000000	-5.000000	-2.000000
25%	1.000000	10.000000	9.000000
50%	3.000000	18.000000	14.000000
75%	5.000000	27.750000	18.000000
max	777.000000	65.000000	4106.000000
std	40.941233	14.656208	224.207674

Change the following < 0 values to NULL

Rationale: Does not make sense from the context of the variables (delay and number are quantities that cannot be negative)



Clip outliers to 97th percentile

Num_Bank_Accounts	10.000000
Num_Credit_Card	10.000000
Interest_Rate	33.000000
Num_of_Loan	9.000000
Delay_from_due_date	57.000000
Num_of_Delayed_Payment	24.130000

Rationale: The context of the variables makes the very high values likely to be an error. 97th percentile seems to have the highest values that still remain reasonable.

GOLD TIER

LABEL STORE

CreditKarma Scorer predicts the probability that a customer will “default” their loan given that it is granted. Thus, we need to define what customers will be considered as “default”.

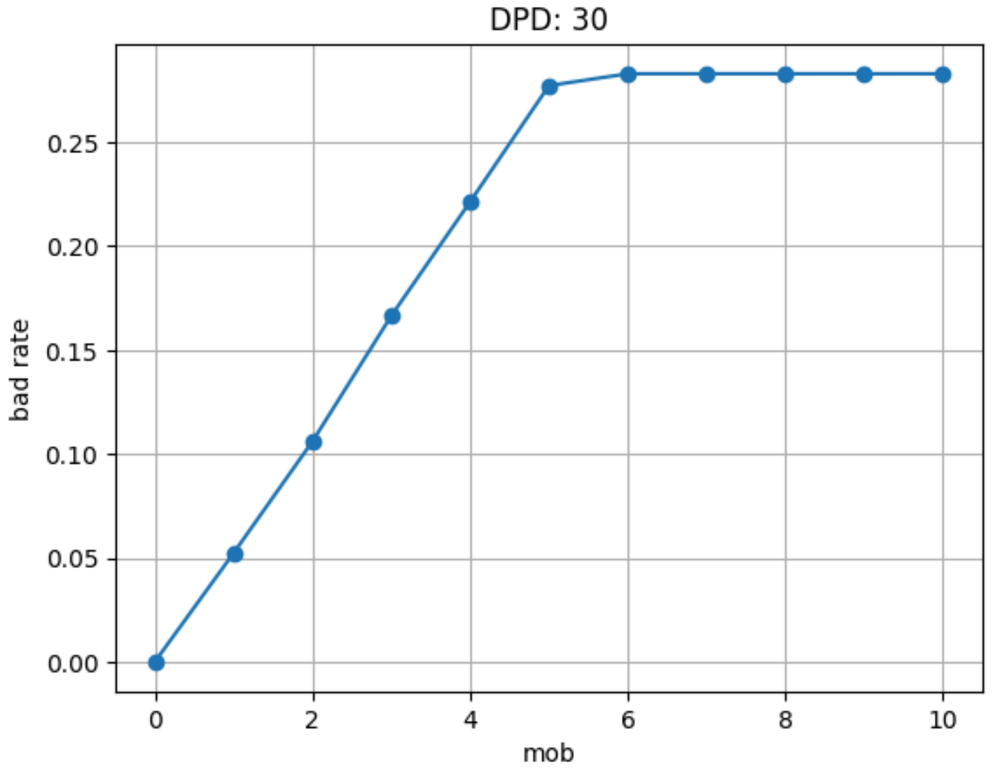
Days Past Due

We want to give some **leniency** to customers before labelling their payment as overdue. Hence, we set a tolerance level of **30 days**.

Bad Rate

A “bad customer” is everyone who has not paid off their loans in the 10th installment. We want to find out how much earlier we can start estimating that a customer will end up as a bad customer by plotting the proportion of “bad customers” who have overdue loans for each month on book.

If a customer is on their **6th month on book** and **still has payments that are 30 days past due**, it is unlikely they will pay them off before the tenure. Hence, these customers are marked as “default”.



GOLD TIER

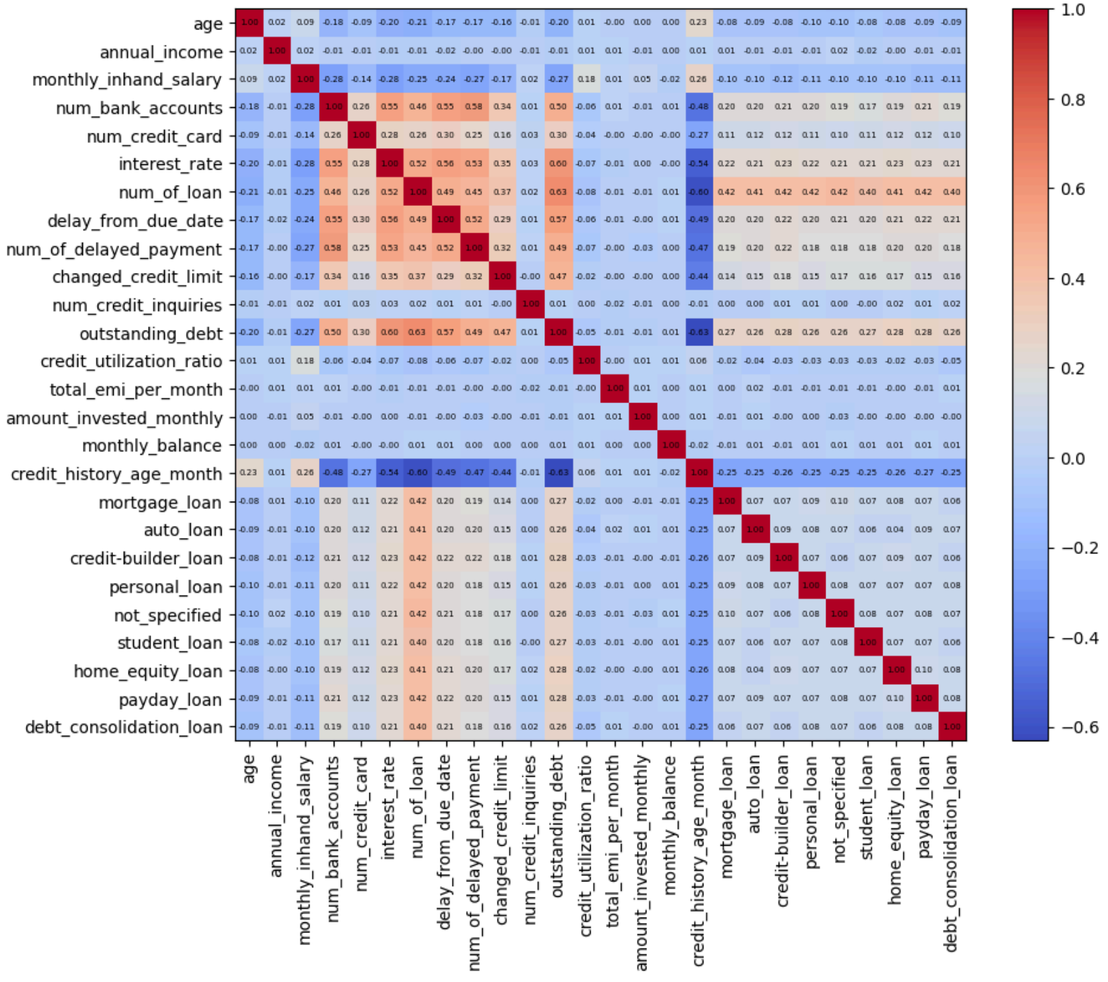
FEATURE STORE

Data from the silver tables are **consolidated into a single feature table** to train the CreditKarma Scorer model.

Feature Selection

The columns of attributes and financials are each meaningful and distinct, so the choice is made to include all the features. However, **credit_history_age_year** and **credit_history_age_month** are combined into one column in terms of months to slightly reduce the dimensionality.

Correlation analysis is also done as CreditKarma Scorer is a logistic regression model, which is sensitive to highly correlated features. However, analysis shows that the features are not heavily correlated, so it is safe to keep them all.

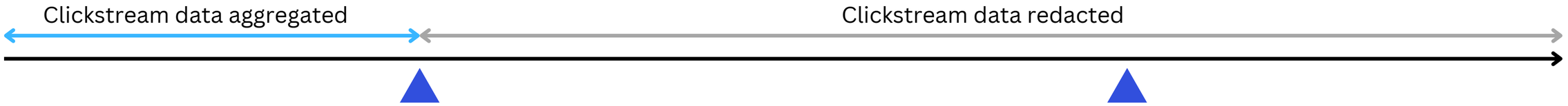


GOLD TIER

CLICKSTREAM PROCESSING

Clickstream data is handled a little differently from attributes and financials, as it captures many snapshots for each user. Therefore, we must take care to **prevent data leakage** and **aggregate data for each user** to capture their behavior in a single record**.**

Preventing data leakage



First installment made

- Attributes and Financials system logs data
- Aggregated clickstream data joined

6th month on book

- Default label decided for customer

Data aggregation

Data is aggregated by taking the **mean** of the clickstream features for each user. PCA was considered to reduce dimensionality, but we found that the clickstream data is not highly correlated enough for the principal components to explain a large portion of the variance.