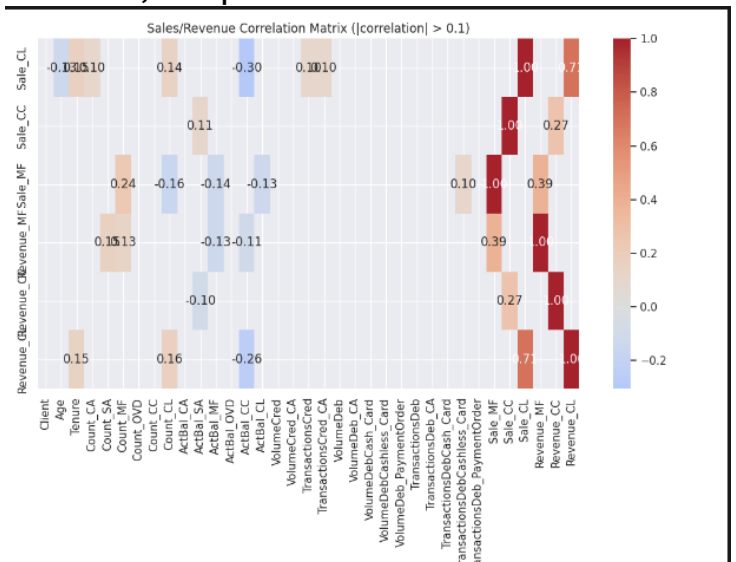
My Solution

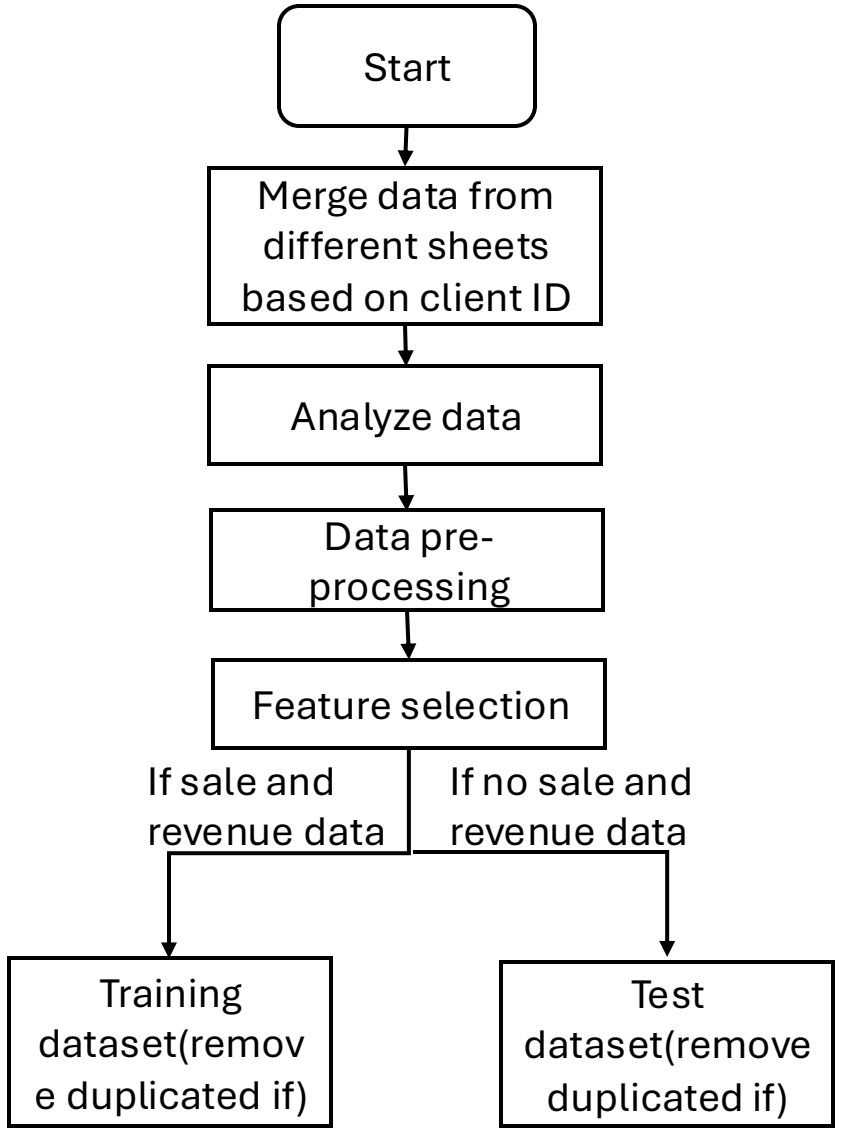
Yuanbin Zhou

7 April 2025

Create Analytical Datasets

- Merge data from different sheets, based on client ID
- [in the next slide]Analyze data and its distribution
- Data pre-processing:
  - Remove anomalies data, for example, age<18, tenure>12*age
  - Use zero, median or KNN to fill missing values depends on concrete cases
  - Use log function to handle right-skewed data
- Feature selection
  - Choose most relevant features according correlation to the targeted feature
  - Filter out highly co-related features
- For data that has sale and revenue data, keep as training data
- For data that don't have sale and revenue data, keep as test dataset
- Remove duplicated data if any

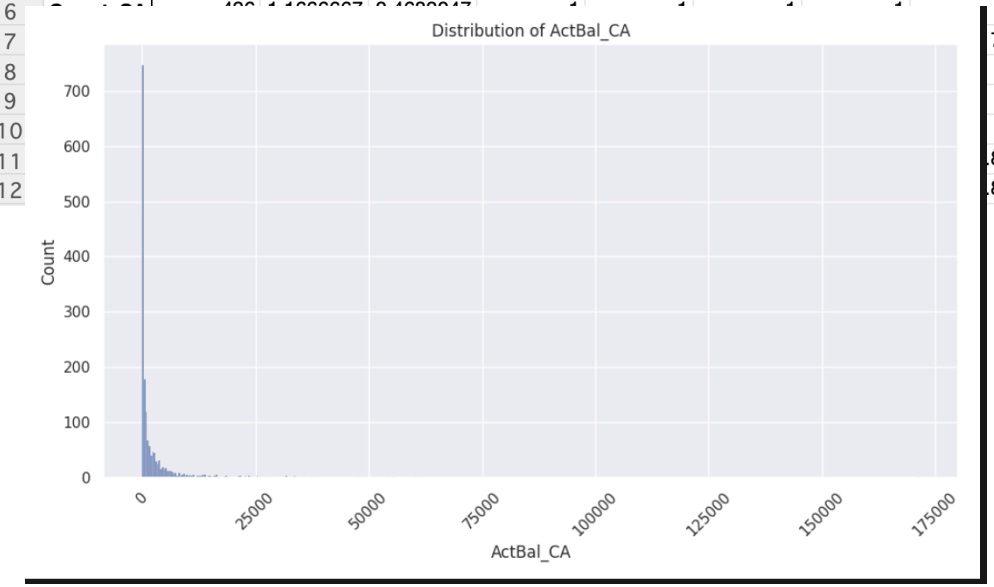


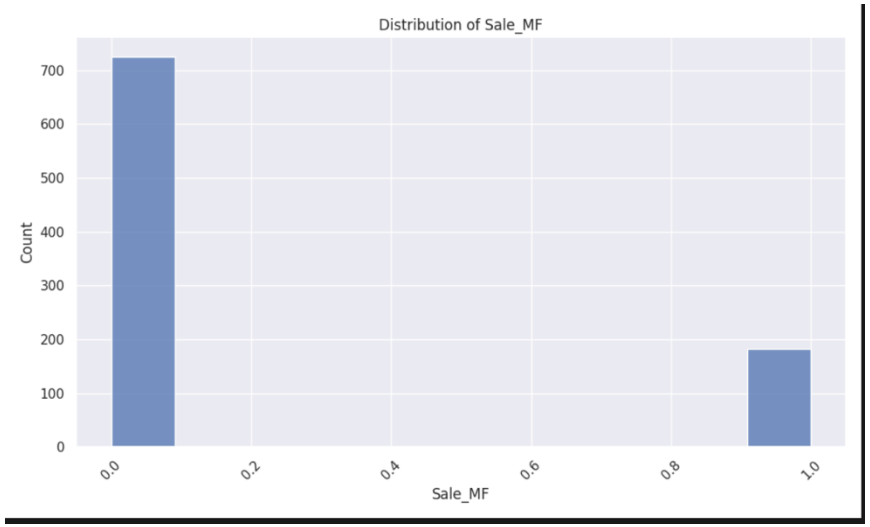


Analyze Data

- Anomaly data: age with zero, tenure>12*age:
- Data missing in the dataset
  - If Count_CA/SA… missing, ActBal_CA/SA… also missing
    - Safe to fill with zero for empty data
  - 5 Count_CC is not empty, but ActBal_CC missing
    - Use median value to fill
  - Data missing in Inflow_Outflow
    - Use KNNImputer
  - Sex only has one missing
- Right-skewed data: balance and transactions data (e.g., ActBal_CA in the second pic)
  - Use log function to handle, may also possible use sqrt, Yeo-Johnson
- Slightly imbalance data: output class Sale_MF
  - Use over-sampling techniques

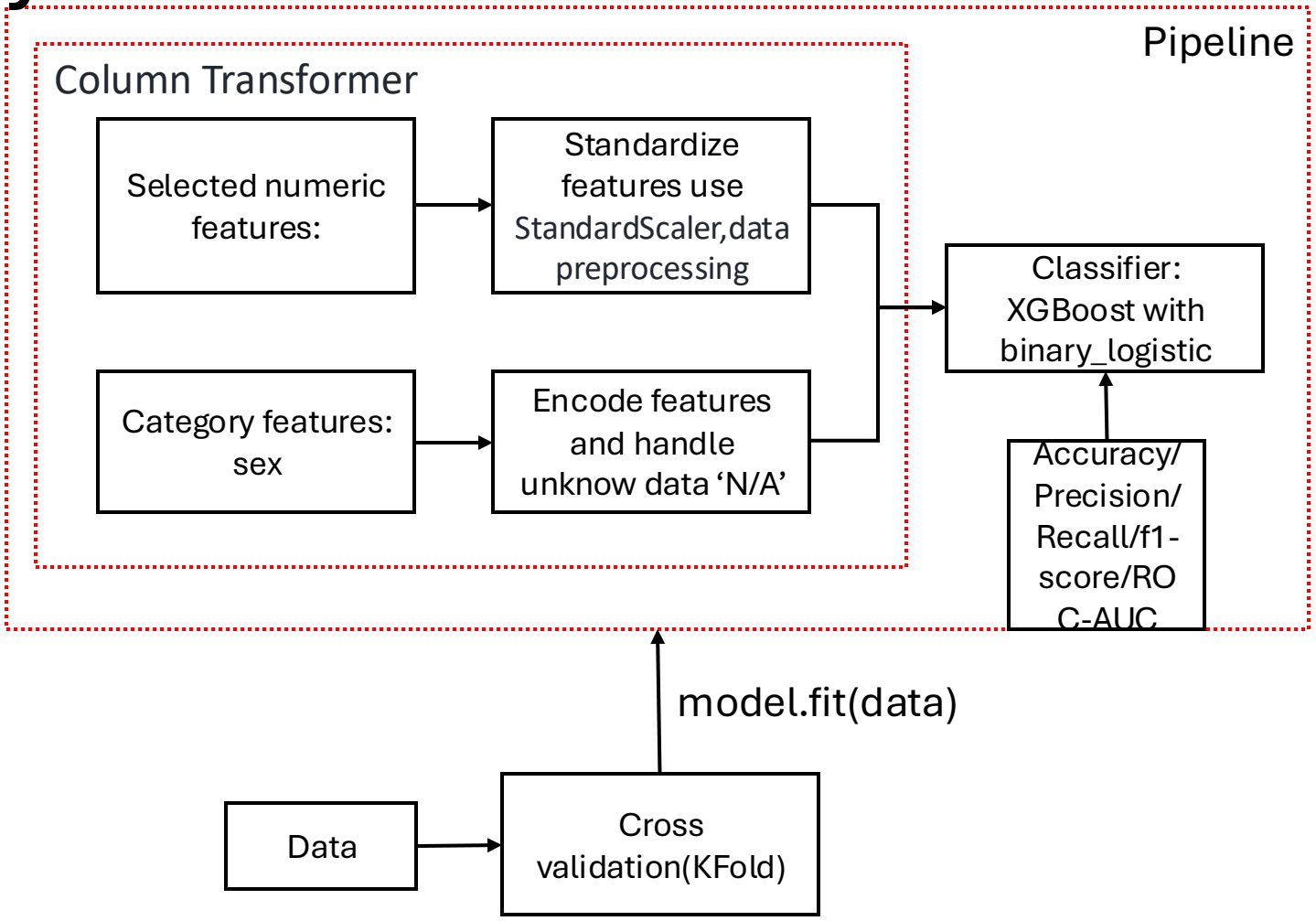
	count	mean	std	min	25%	50%	75%	max
Client	1615	808	466.35466	1	404.5	808	1211.5	1615
Age	1615	42.848916	18.550529	0	29	41	57	97
Tenure	1615	101.33994	64.917297	0	44	97	151	273
Count_CA	1615	1.0786378	0.3330355	1	1	1	1	4





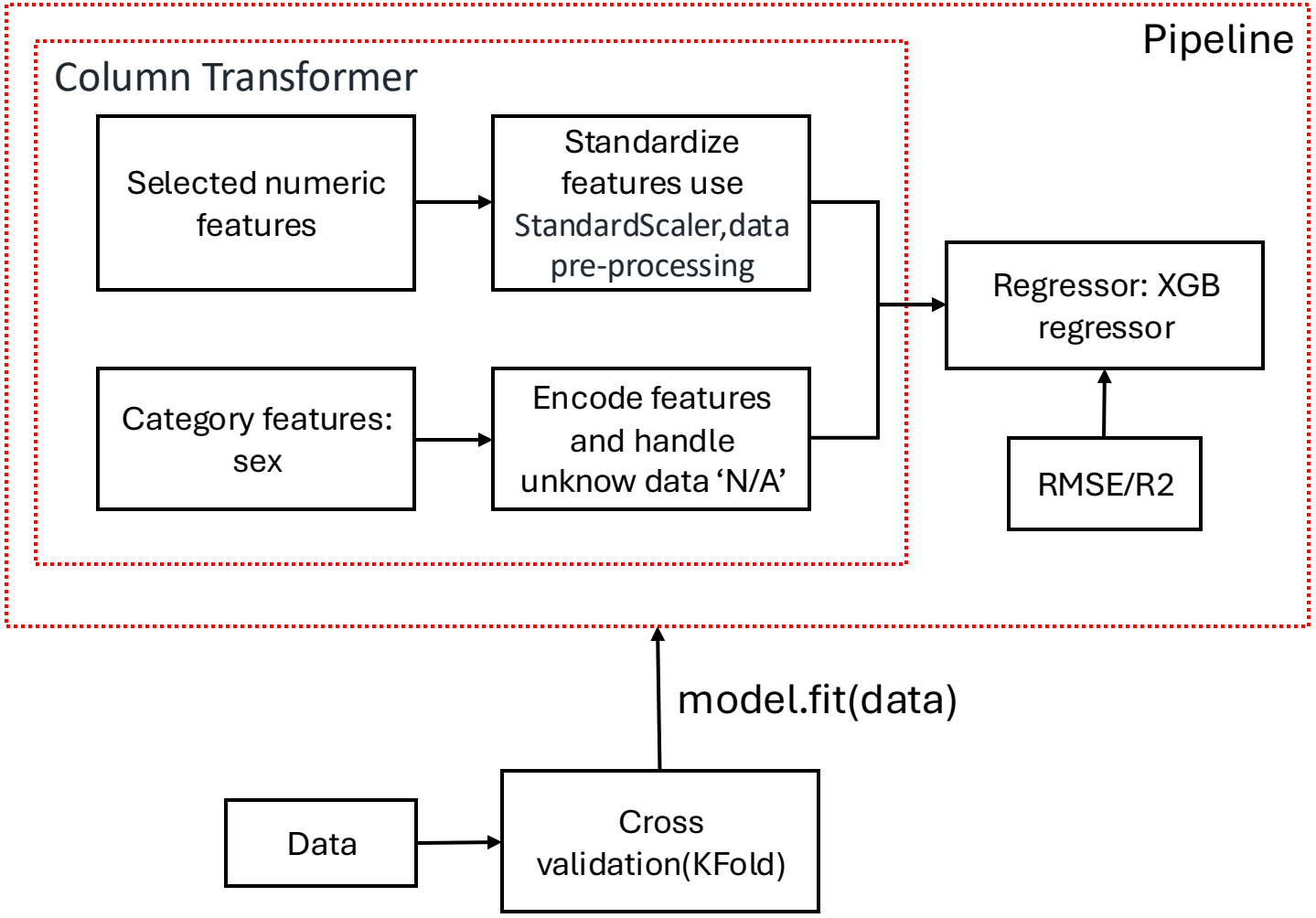
Develop Propensity Models

- Include numeric and category features
- Numeric features: to standardize features, using StandardScaler method,
  - Use KNN/zero to fill empty data
- Category features: encode as a one-hot numeric array
- Different features concatenated into a single feature space use columntransformer
- Apply the pipeline to preprocess the data and with a final classifier
- Handle imbalance data use over-sampling SMOTE
- Use cross validation to split training and validation data set
- Classifier use XGBoost with binary_logistic



Optimize Targeting Strategy

- Pipeline is similar to the previous
- Use regression model instead of classification model
- Evaluation metrics: RMSE



Maximize Revenue

- Calculate expected revenue=likelihood*revenue
- For each client, get the maximum expected revenue from either CC, CL, MF
- Based on the above data, get top 100 clients that maximize revenue
- May also possible to use ILP to solve the problem if constraints put to the number of each type of offer.