Service Payment Call Predictive Modeling

07/30/2020

Hwanpyo Kim, Ph.D.

Overview

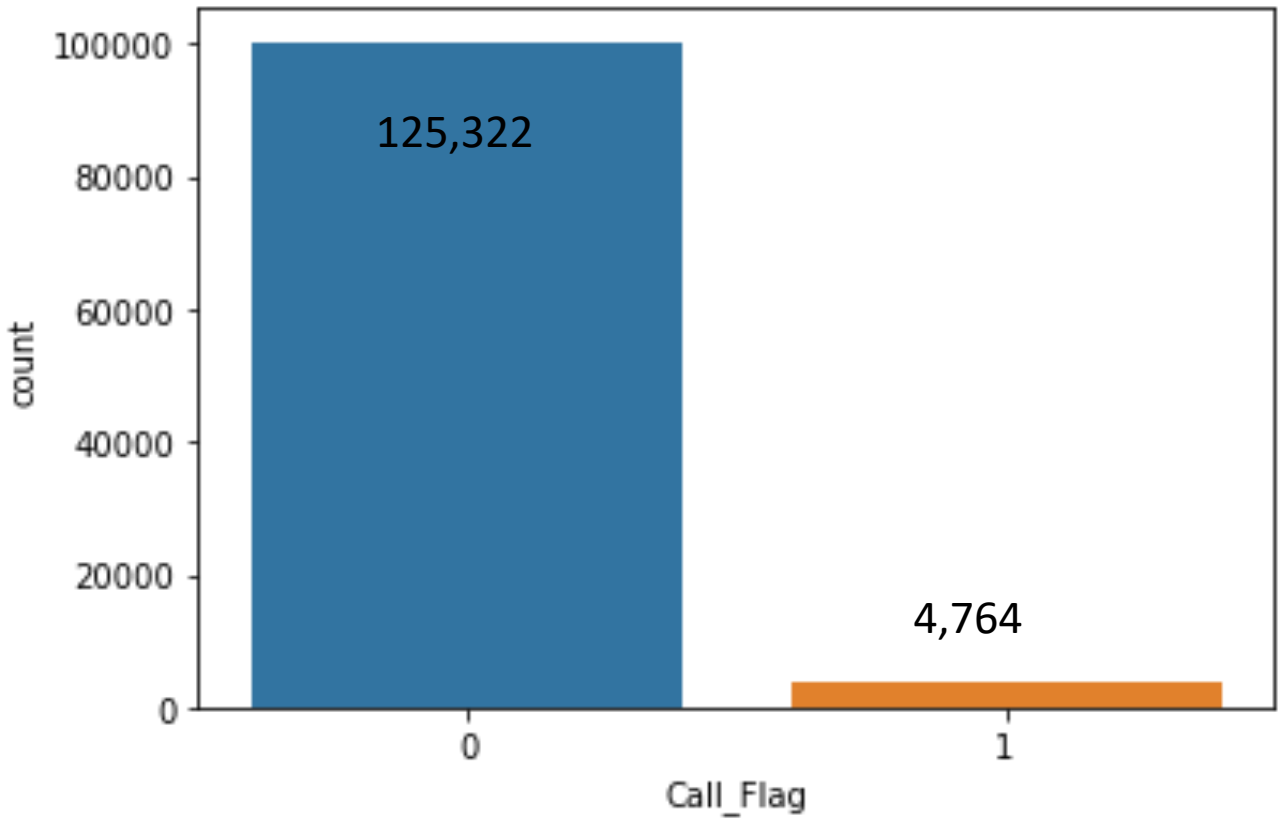
- Introduction
- Data Analysis – Imbalanced dataset
- Predictive Modeling Framework
- Conclusion

Introduction

- Objective…
  - We want to predict the likelihood that each policyholder will make a service payment call in the next 5 days.
  - It will help to select people to receive a pre-emptive email message designed to encourage them to pay through GEICO's self-service channels
- What we have…
  - Data on customers who have had a bill due in the next 5 days and whether they made a service payment (130,086 records with 29 variables)
- Scope of the presentation
  - Exploratory analysis/Data pre-processing
  - Predictive modeling with Imbalanced target
  - Performance evaluation

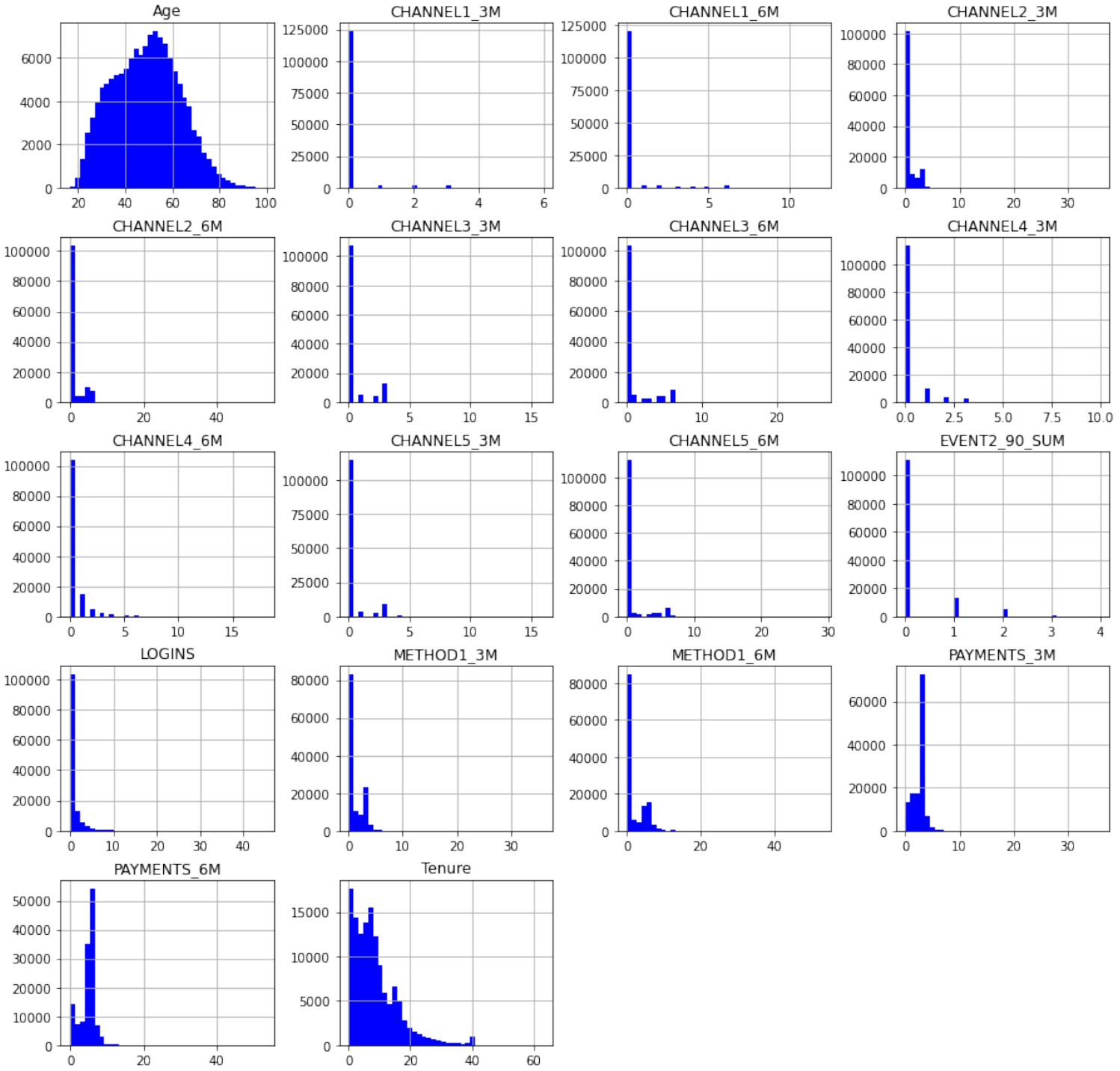
Review on Data

- 130,086 records and 29 variables
- Target Variable: “Call_Flag”
  - Imbalanced records (26:1)
- 20 Numerical Variables: (continuous/discrete/date: dropped)
- 9 Categorical Data: (nominal)
- Missing values with 809 rows within the specific fields



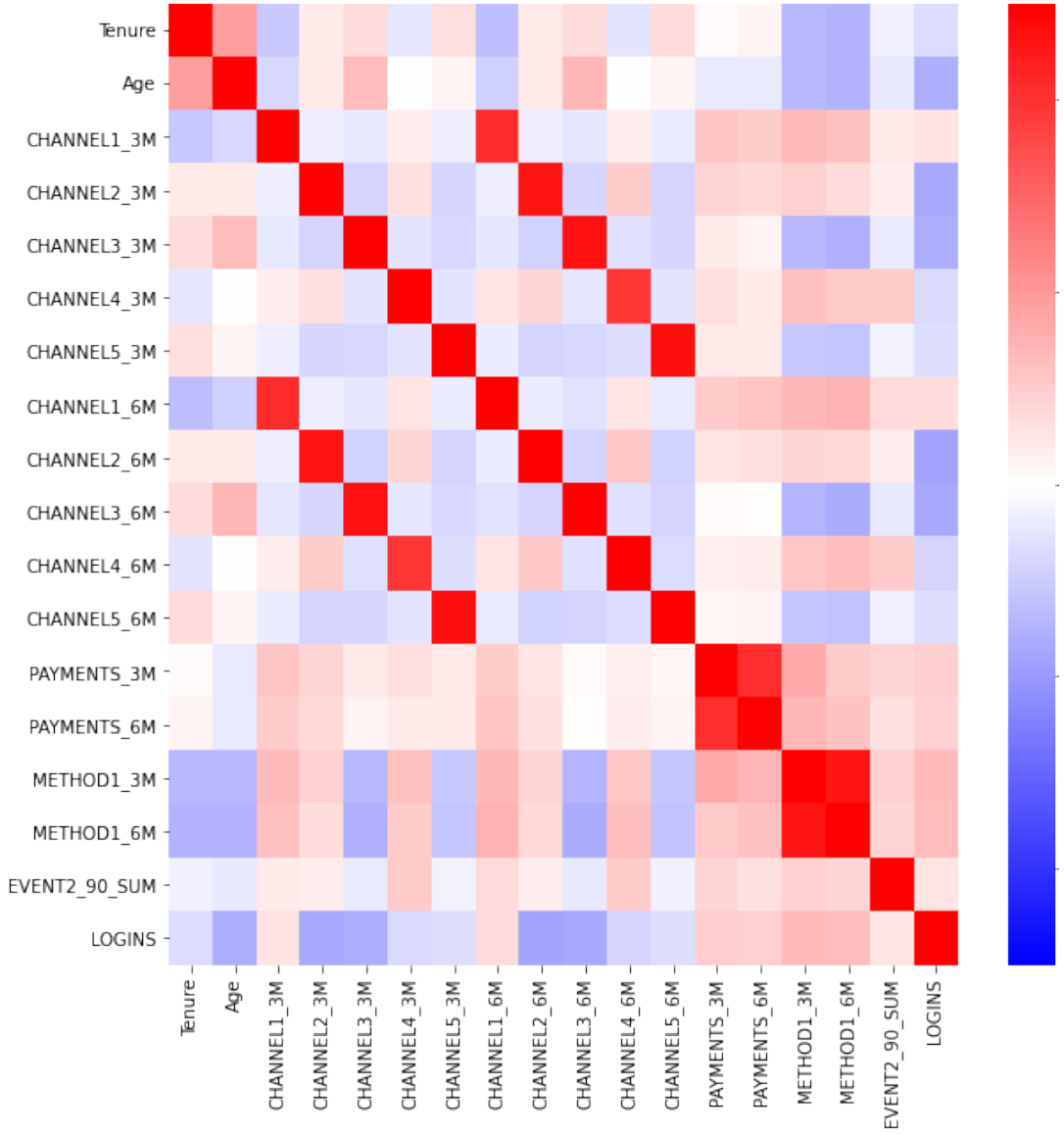
Distribution of Numerical Variables

- Strongly Skewed distributions of continuous & discrete variables
- “Tenure” is transformed into “Log_Tenure”.



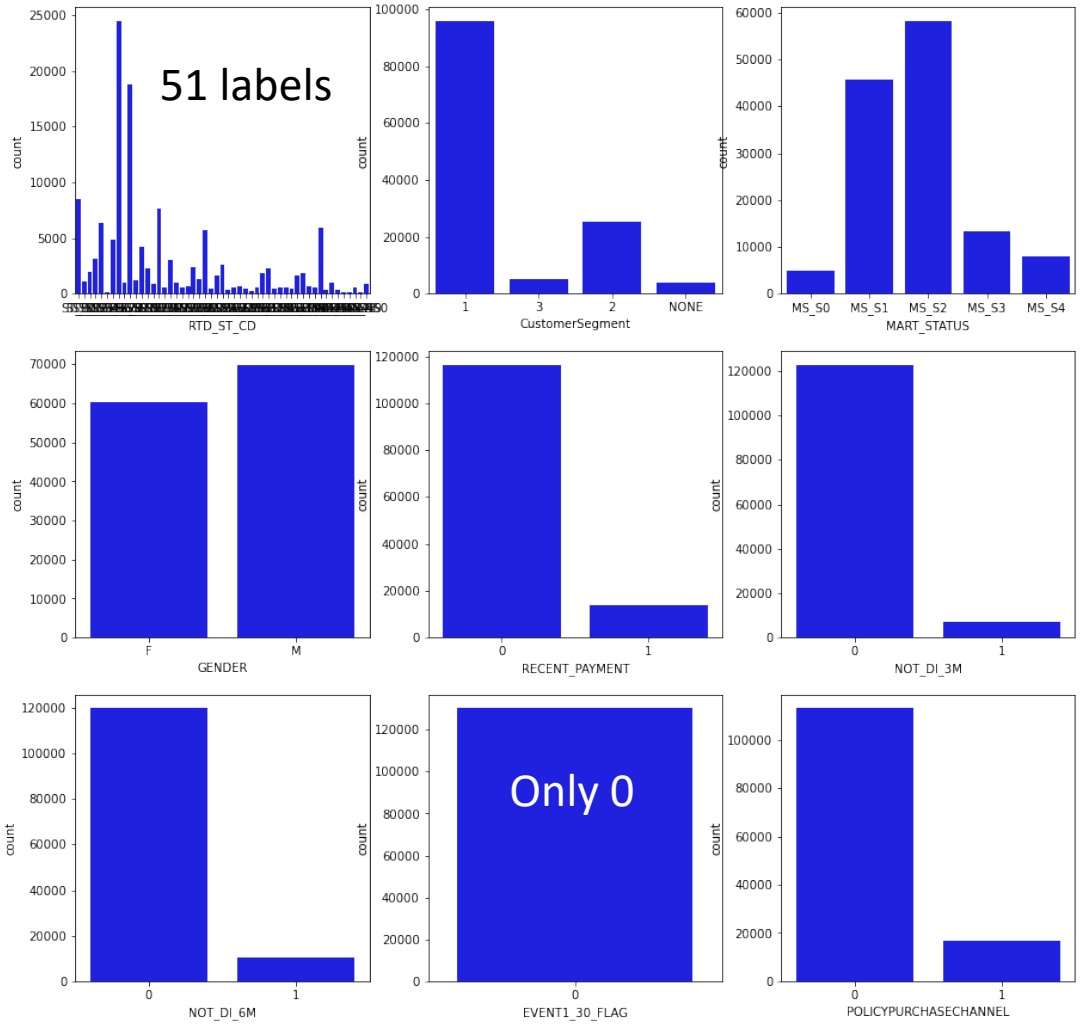
Correlation of Numerical Variables

- Strong Correlations between “CHANNELx_3M” &“CHANNELx6M”, “METHOD1_3M” & ”METHOD1_6M”, “PAYMENTS_3M” & “PAYMENTS_6M”
- Spearman correlation (ρ): nonparametric measure of rank correlation. (monotonic relationship)
  - Assumption of Pearson correlation (linear relationship) are violated: skewed distribution



Categorical Variables

- “EVENT1_30_FLAG”, all zero values, should be dropped
- There are lots of labels in “RTD_ST_CD”: We need feature selection approaches!
- One-Hot encoding

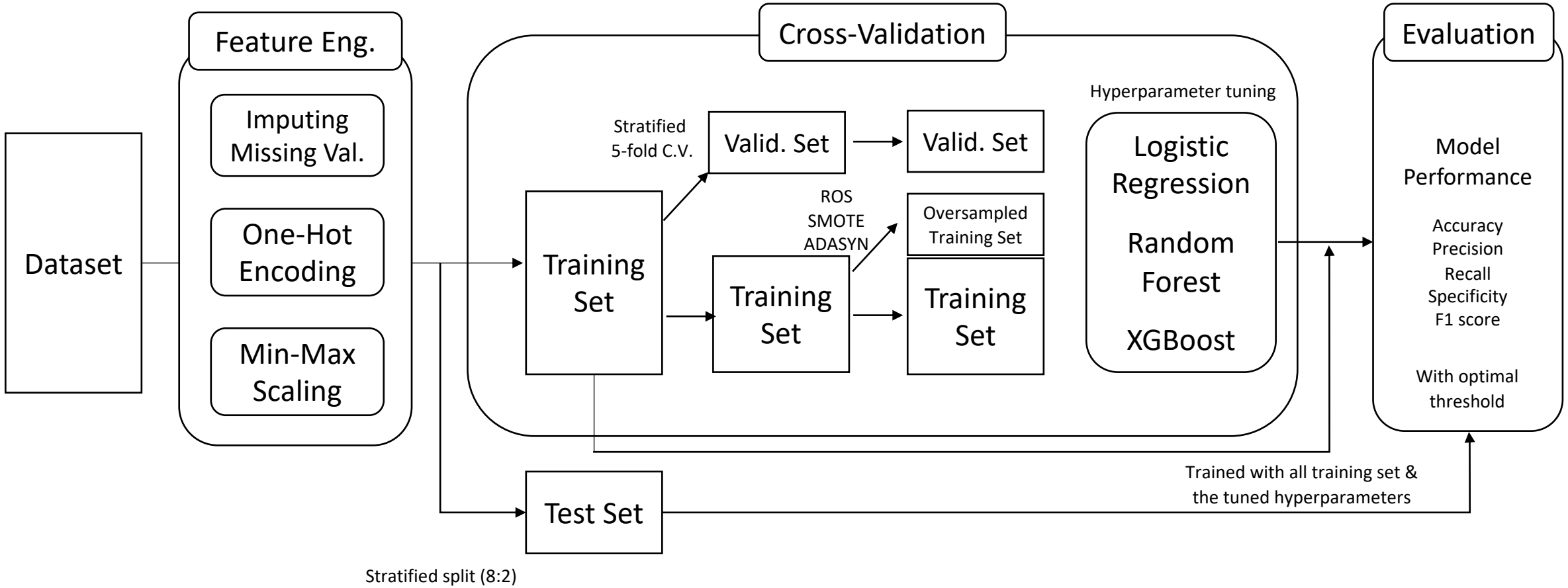


Feature Engineering

- Imputing Missing Values
  - 809 missing rows in the following colums, "CHANNELx_6M", "METHOD1_6M", "RECENT_PAYMENT", "PAYMENTS_6M", simultaneously.
  - All of missing columns are related with payment.
  - "NOT_DI_xM", “CHANNELx_3M”, “METHOD1_3M”, “PAYMENTS-3M” are almost zero in the corresponding instances.
  - I assume that every NaN values are zeros.
- One-Hot Encoding
- Min-Max Scaling

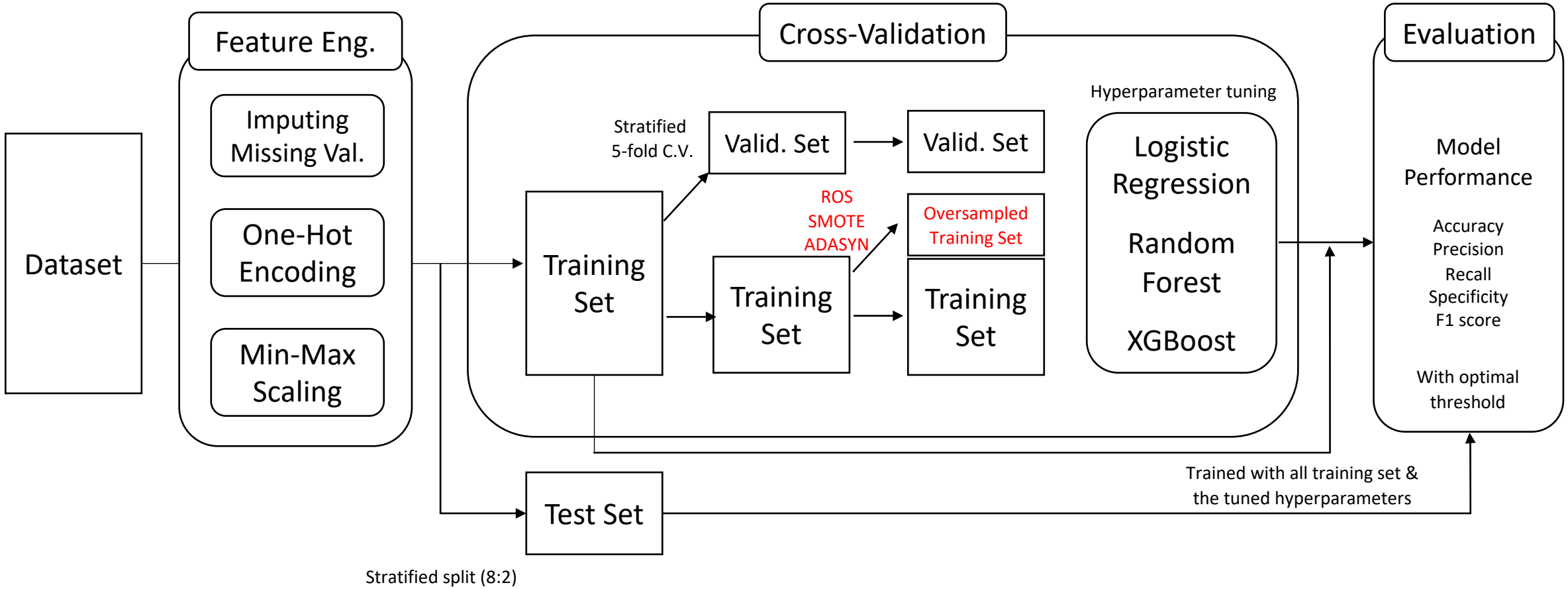
ML Modeling Workflow

with Imbalanced dataset



ML Modeling Workflow

with Imbalanced dataset



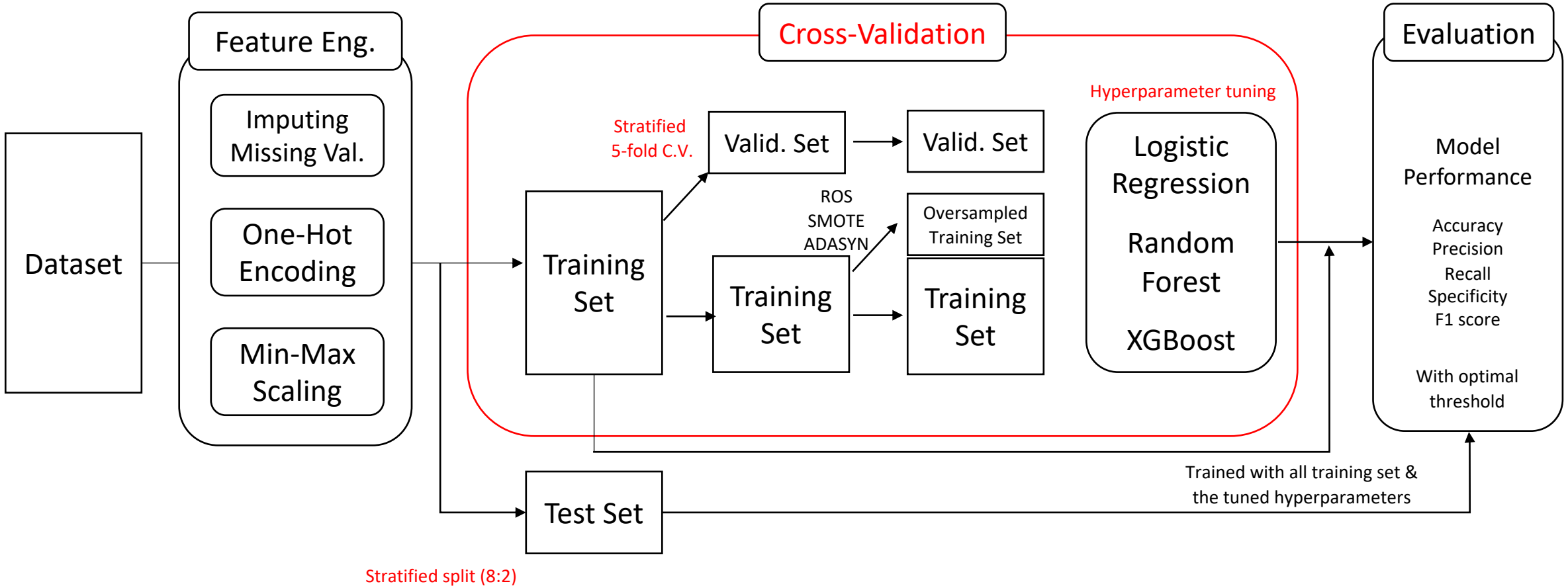
Over-sampling

- Resampling techniques to balance classes are considered
  - Random UnderSampling: eliminate majority class examples. (Information loss)
  - Random OverSampling: replicate instances in the minority class randomly.
    - Cons: Overfitting with replication of the minority events
  - Synthetic OverSampling: generate new samples in by interpolation to prevent overfitting
    - SMOTE*: generating samples from a drawn line between the examples in the feature space. (any distinction between easy and hard samples to be classified)
    - ADASYS**: generating samples next to the original samples which are wrongly classified (inversely proportional to the density)

* N. V. Chawla, K. W. Bowyer, L. O.Hall, W. P. Kegelmeyer, “SMOTE: synthetic minority over-sampling technique,” Journal of artificial intelligence research, 16, 321-357, 2002.
** He, Haibo, Yang Bai, Edwardo A. Garcia, and Shutao Li. “ADASYN: Adaptive synthetic sampling approach for imbalanced learning,” In IEEE International Joint Conference on Neural Networks (IEEE World Congress on Computational Intelligence), pp. 1322-1328, 2008.

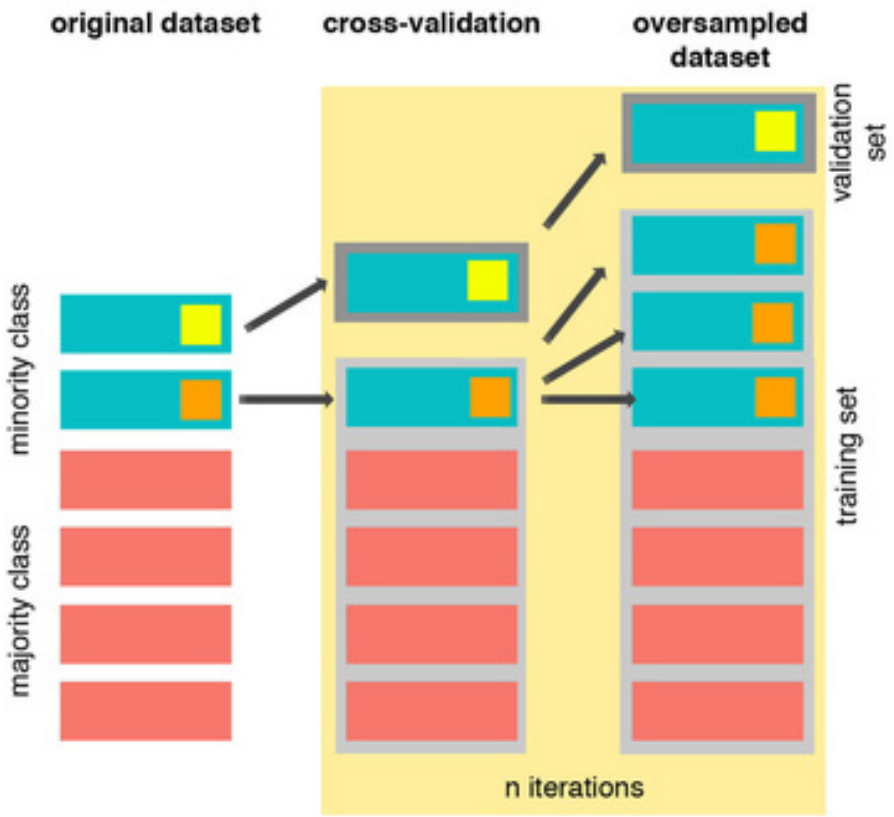
ML Modeling Workflow

with Imbalanced dataset



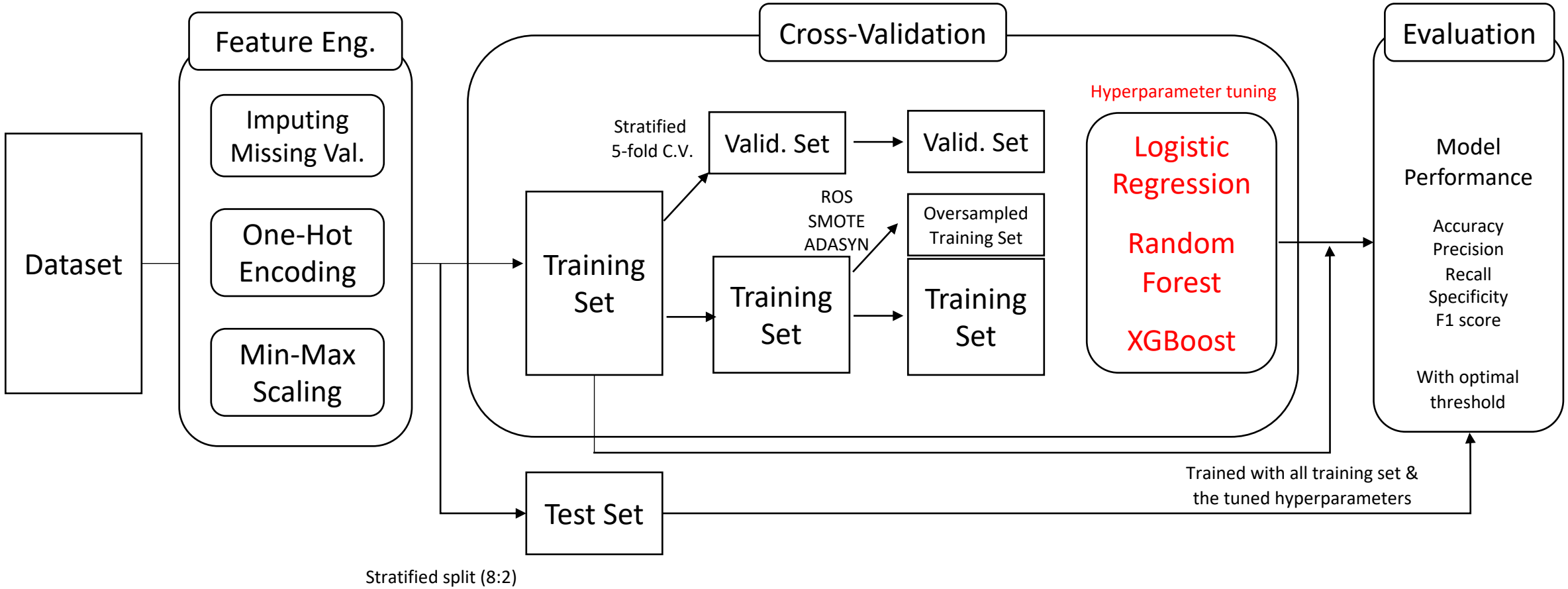
Train/Test split and Cross-Validation with oversampling

- Stratified train/test Split and k-fold Cross-Validation
  - enforce class distribution in training/test set and each fold in C.V identically.
- Oversampling within Cross-validation
  - We need oversampling on training set at every iteration in C.V.
  - To prevent overfitting from replicated/synthetic samples in valid. set
  - “imblearn.pipeline.Pipeline”
- Score-metric in C.V.
  - Average precision score (≈ Precision-Recall AUC)



ML Modeling Workflow

with Imbalanced dataset

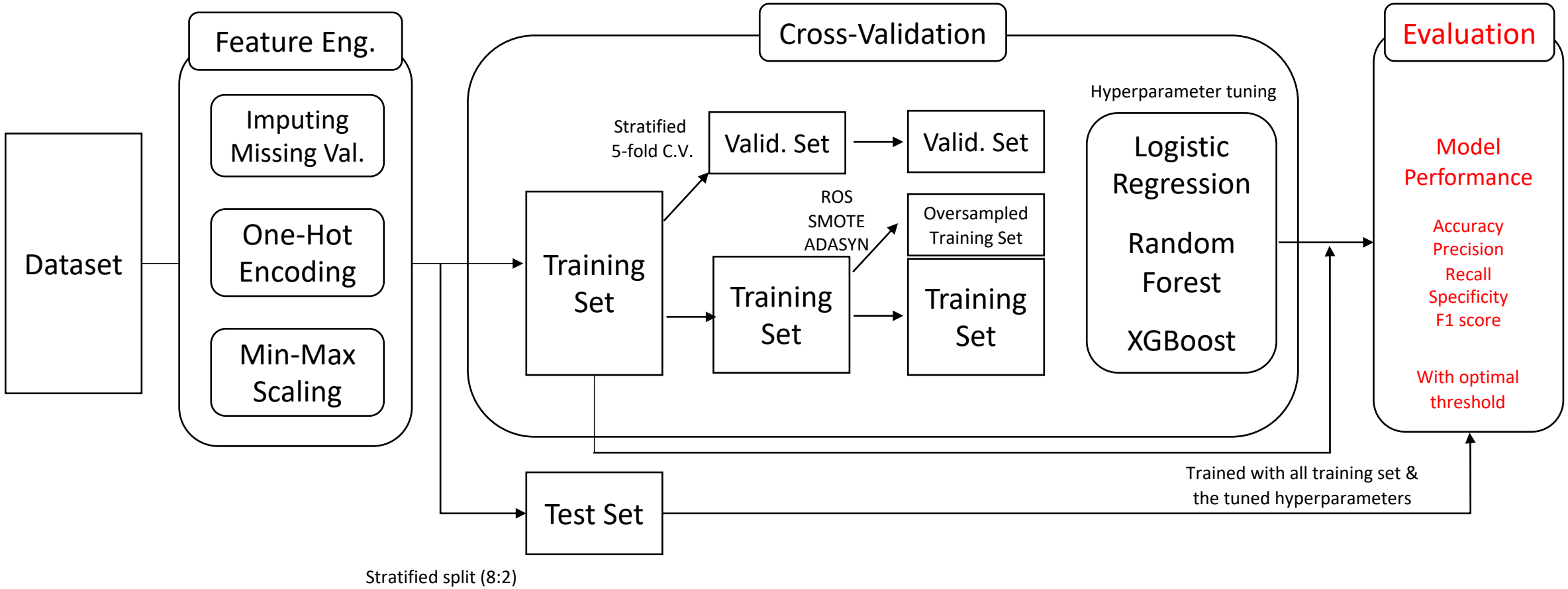


Classification Models

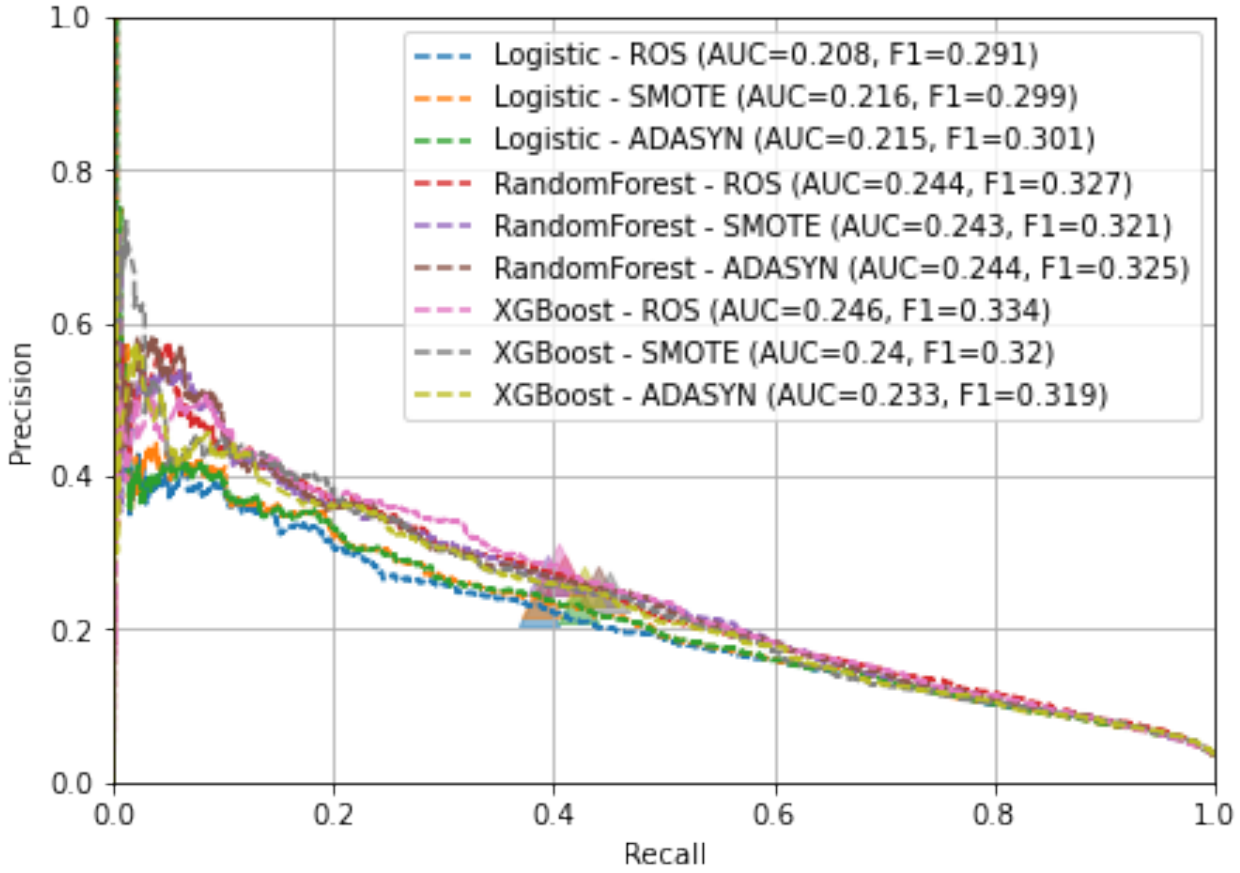
- Logistic Regression:
  - A simple GLM for classification with L1 regularization to feature selection
  - Hyperparameters: Inverse of regularization strength “C”
  - Coefficient of features
- Ensemble Algorithms:
  - XGBoost: Gradient Boosting with advanced merits
  - Random Forest: Collection of random trees with Bagging
  - Hyperparmeters: number of trees “n_estimators”, maximum depth of tree “max_depth”, etc.
  - Variable importance (Gain/avg. decrease in Gini impurity)

ML Modeling Workflow

with Imbalanced dataset



Model Performances (1)



- Precision-Recall Curves
- Random Forest & XGBoost superior than Logistic Regression w.r.t. PR-AUC

Model Performance (2)

- Scores at optimal thresholds to maximize F1 Score

Model	O.S.	Accuracy	Precision	Recall	F1 score
Logistic	ROS	0.931	0.233	0.387	0.291
	SMOTE	0.933	0.244	0.388	0.299
	ADASYN	0.928	0.233	0.423	0.301
Random Forest	ROS	0.963	0.272	0.408	0.327
	SMOTE	0.939	0.272	0.393	0.321
	ADASYN	0.933	0.258	0.440	0.325
XGBoost	ROS	0.941	0.285	0.405	0.334
	SMOTE	0.930	0.248	0.450	0.320
	ADASYN	0.933	0.255	0.426	0.319

Variance Importance

- “CHANNEL4”, “CHANNEL2”, “PAYMENTS” “LOGINS” are high-ranked among the models.
- These variables are related to the possibility of a customer that make a service payment call.

- Future works: feature dimensional reduction