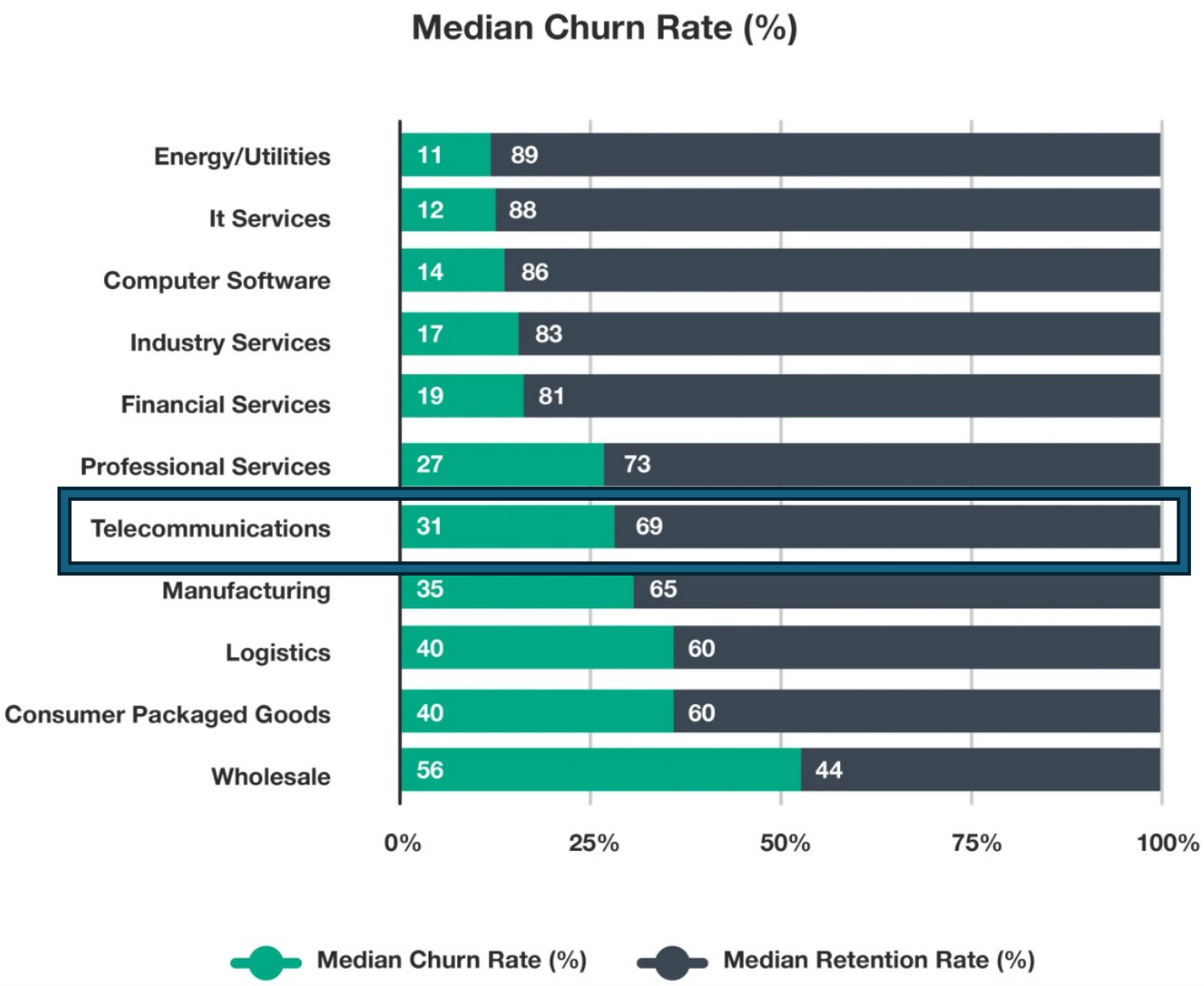
Introduction to Machine Learning

Group 5

Aman Sharma, Muhammad Ibrahim, Téa McCormack

Table of Contents

Business Problem



ARPU (Average Revenue Per Unit) = $60 USD

What's the Average Churn Rate by Industry?
Virgin Media Q2 sets up for 2024 execution with focused investments in Q1

Business Problem

- **WHAT:**
  - Customers leaving
  - Customers getting discounts
- **WHY:**
  - Save Customer?
    Cost of Acquisition > Cost of Retention
  - Save Discounts?
    This can add up if everyone gets it
- **GOAL:**
  - Predict potential churns

Understanding the Dataset

Customer Demographic	Account Information	Services
Gender (character)	Tenure (integer)	Phone Service (character)
Senior Citizen (binary – 0/1)	Contract (character)	Multiple Lines (character)
Partner (character)	Paperless Billing (character)	Internet Service (character)
Dependents (character)	Payment Method (character)	Online Security (character)
	Monthly Charges (numeric)	Online Backup (character)
	Total Charges (numeric)	Device Protection (character)
		Tech Support (character)
		Streaming TV (character)
		Streaming Movies (character)

- **DATA:**
  - Consists of 20 features + 1 target variable - Churn
  - Evaluates 7043 customers

```
Churn
       No       Yes
0.7346301 0.2653699
```

EDA & Feature Engineering

3 Step Process

1. **Plot raw data relationship against target**

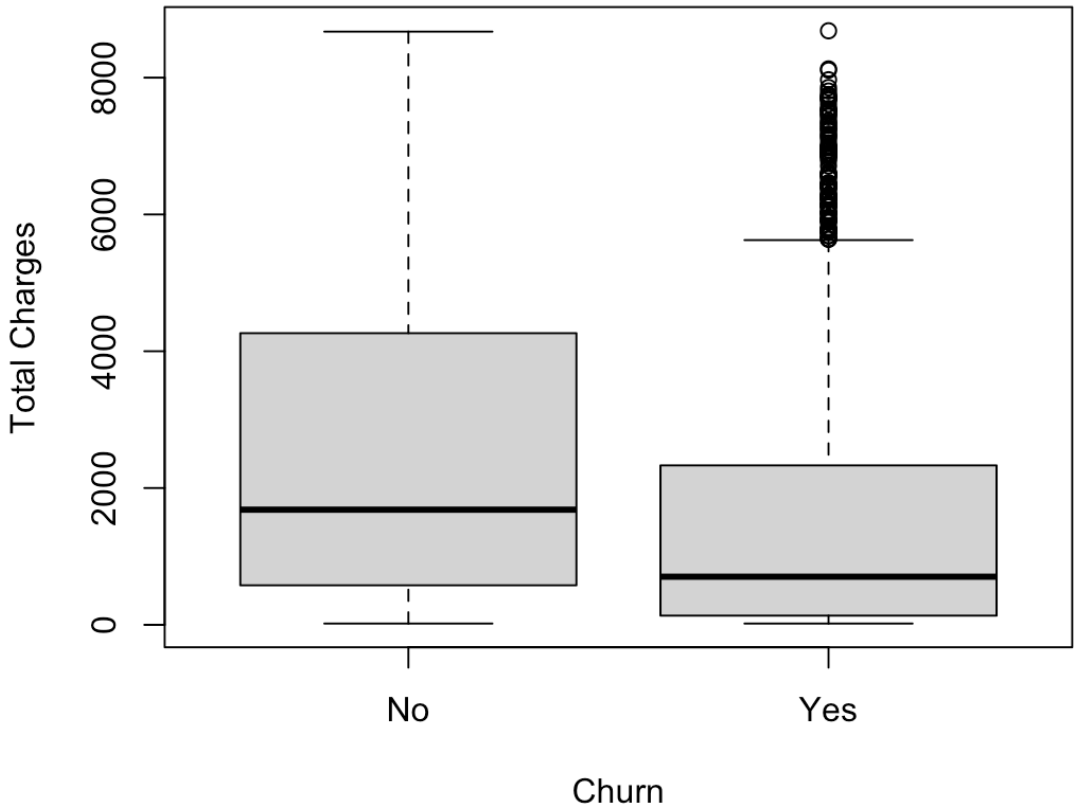
2. Data Wrangling

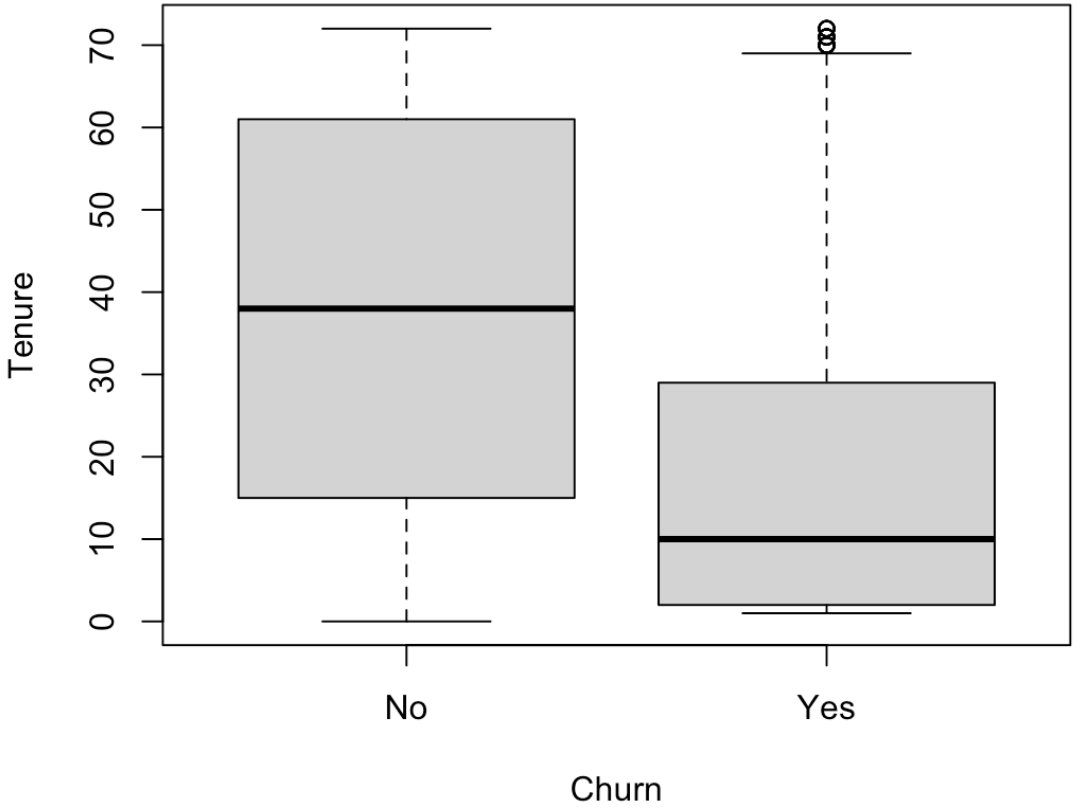
- Imputed null values for total charges
- Dropped customer ID

3. Feature Engineering

- Add_on_services & Charge_ratio
- Assessing feature impact using AUC improvement

EDA & Feature Engineering





EDA & Feature Engineering

INTERNET SERVICE

InternetService	No	Yes
<chr>	<dbl>	<dbl>
DSL	0.810	0.190
Fiber optic	0.581	0.419
No	0.926	0.0740

STREAMING TV

StreamingTV	No	Yes
<chr>	<dbl>	<dbl>
No	0.665	0.335
No internet service	0.926	0.0740
Yes	0.699	0.301

EDA & Feature Engineering

All features (no egg)

```
Resampling results:

  Accuracy   Kappa      AUC_ROC    TPR        FPR        logLoss
  0.8041168  0.46841    0.8450455  .8975845   0.4545145  0.4176579
```

All features (+ Ratio)

```
Resampling results:

  Accuracy   Kappa      AUC_ROC    R          FPR        logLoss
  0.8078436  0.47393    0.8504859  9048309    0.4605324  0.4100031
```

All features (+ Add on)

```
Resampling results:

  Accuracy   Kappa      AUC_ROC    TPR        FPR        logLoss
  0.8053576  0.47065    0.8448905  .8992754   0.4545324  0.417748
```

All features (+ Add on + Ratio)

```
Resampling results:

  Accuracy   Kappa      AUC_ROC    R          FPR        logLoss
  0.8103271  0.48040    0.8511265  9070048    0.4572081  0.4091587
```

- Feature assessment done on Logistic Regression

Model (1/4) : Logistic Regression

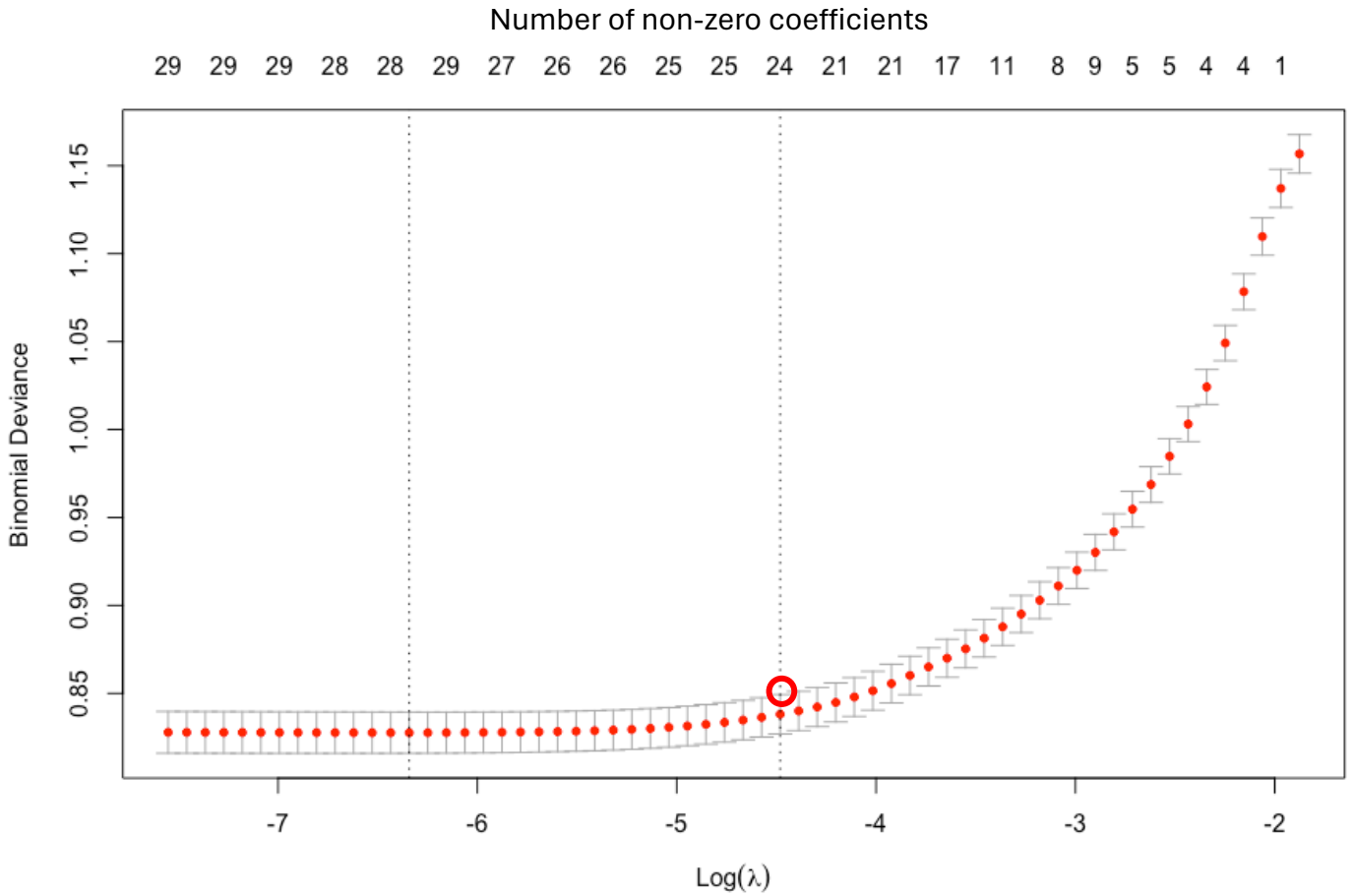
$$Cost\ Function = Logloss + \lambda \sum_{j=1}^{p} |\beta_j|$$

λ = 0.0113

- Lambda based on one SE rule

Confusion Matrix on CV Folds at default 50% threshold

Prediction	Reference No	Reference Yes
No	66.6	12.1
Yes	6.8	14.4

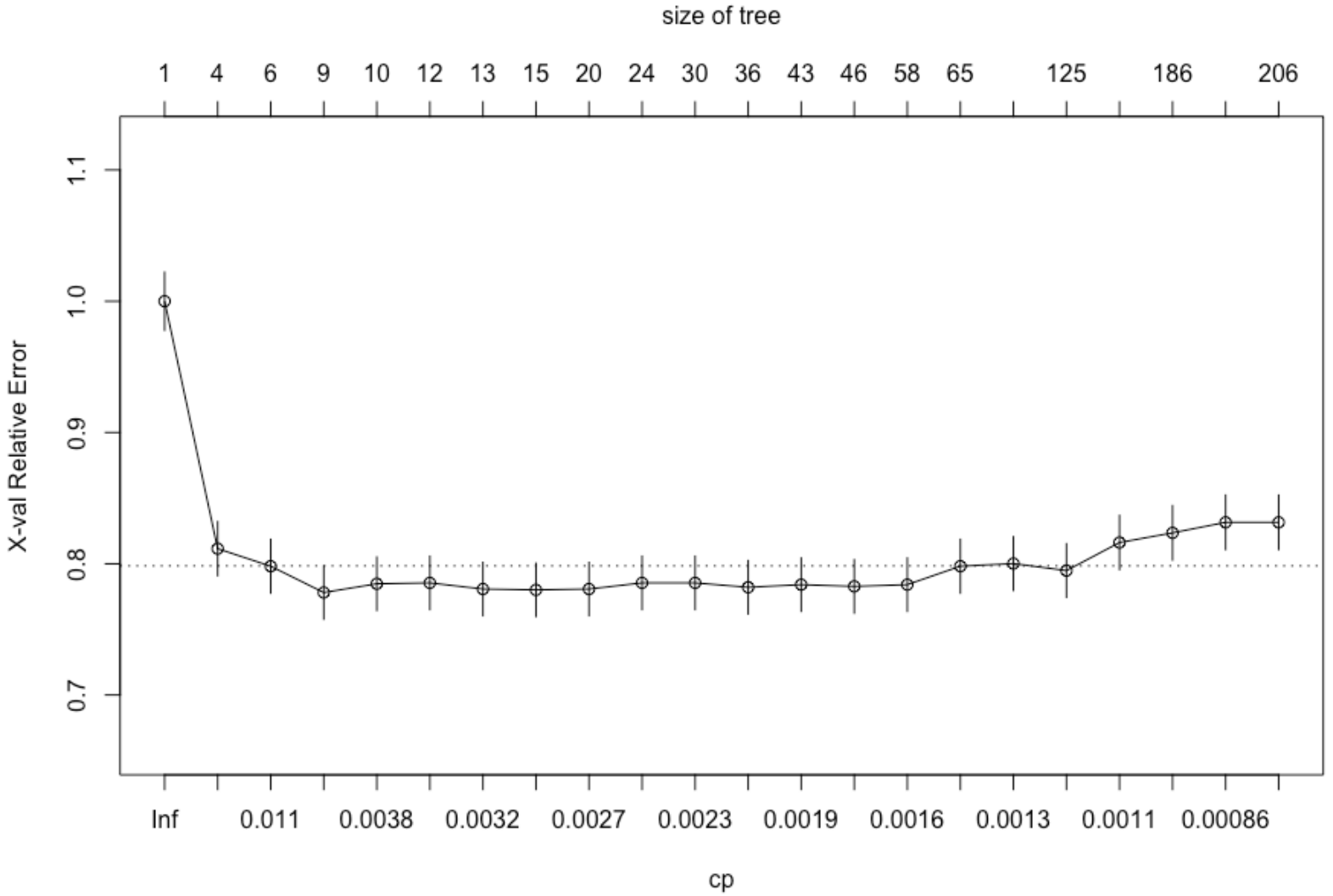


Model (2/4) : Decision Trees

$size\ of\ tree$ = 6

Confusion Matrix on CV Folds at default 50% threshold

	Reference	
Prediction	Yes	No
Yes	15.6	5.5
No	10.9	67.9

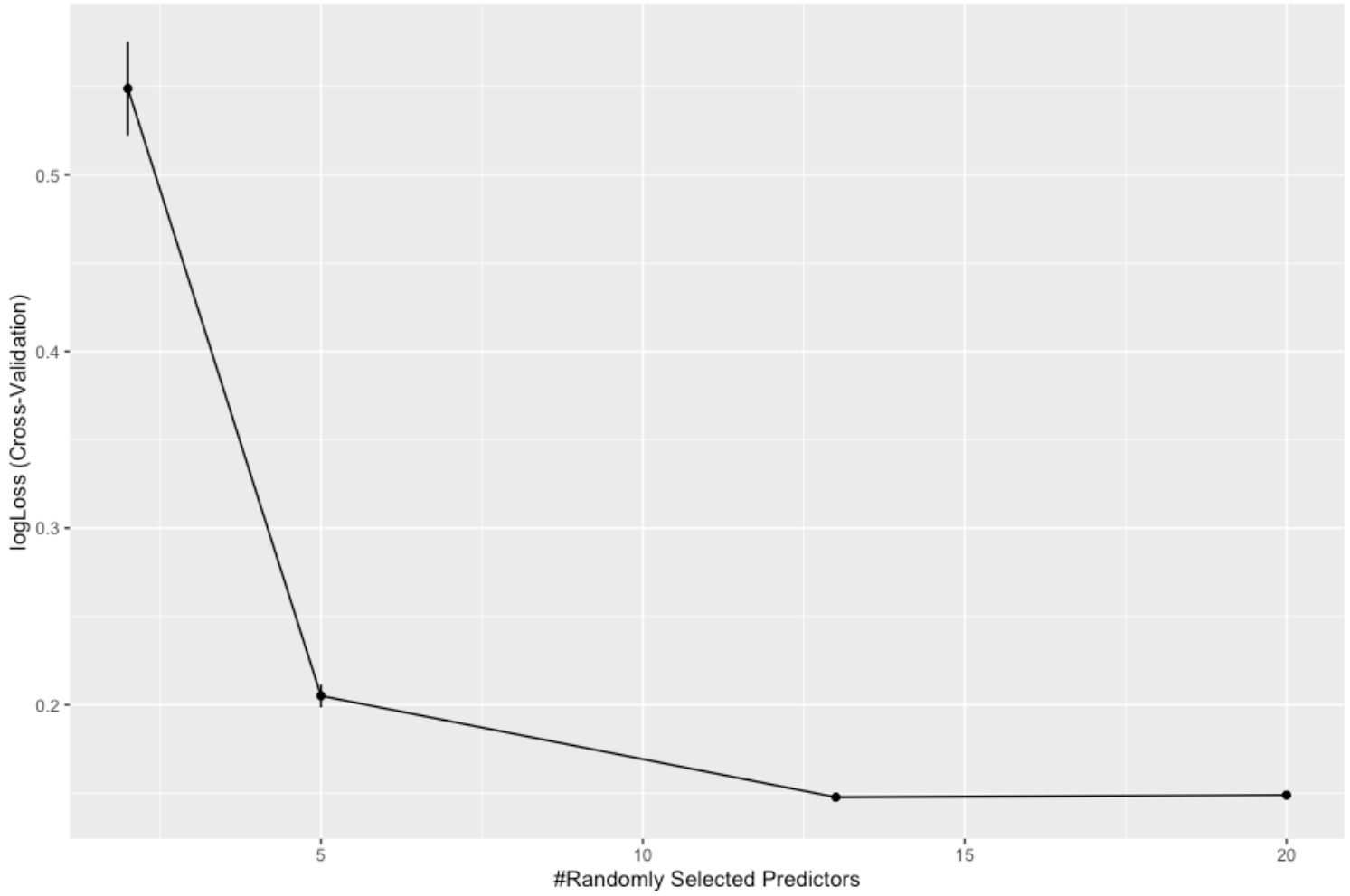


Model (3/4) : Random Forests

$mtry$ = 13

Confusion Matrix on CV Folds

	Reference	
Prediction	No	Yes
No	72.6	1.4
Yes	0.9	25.1



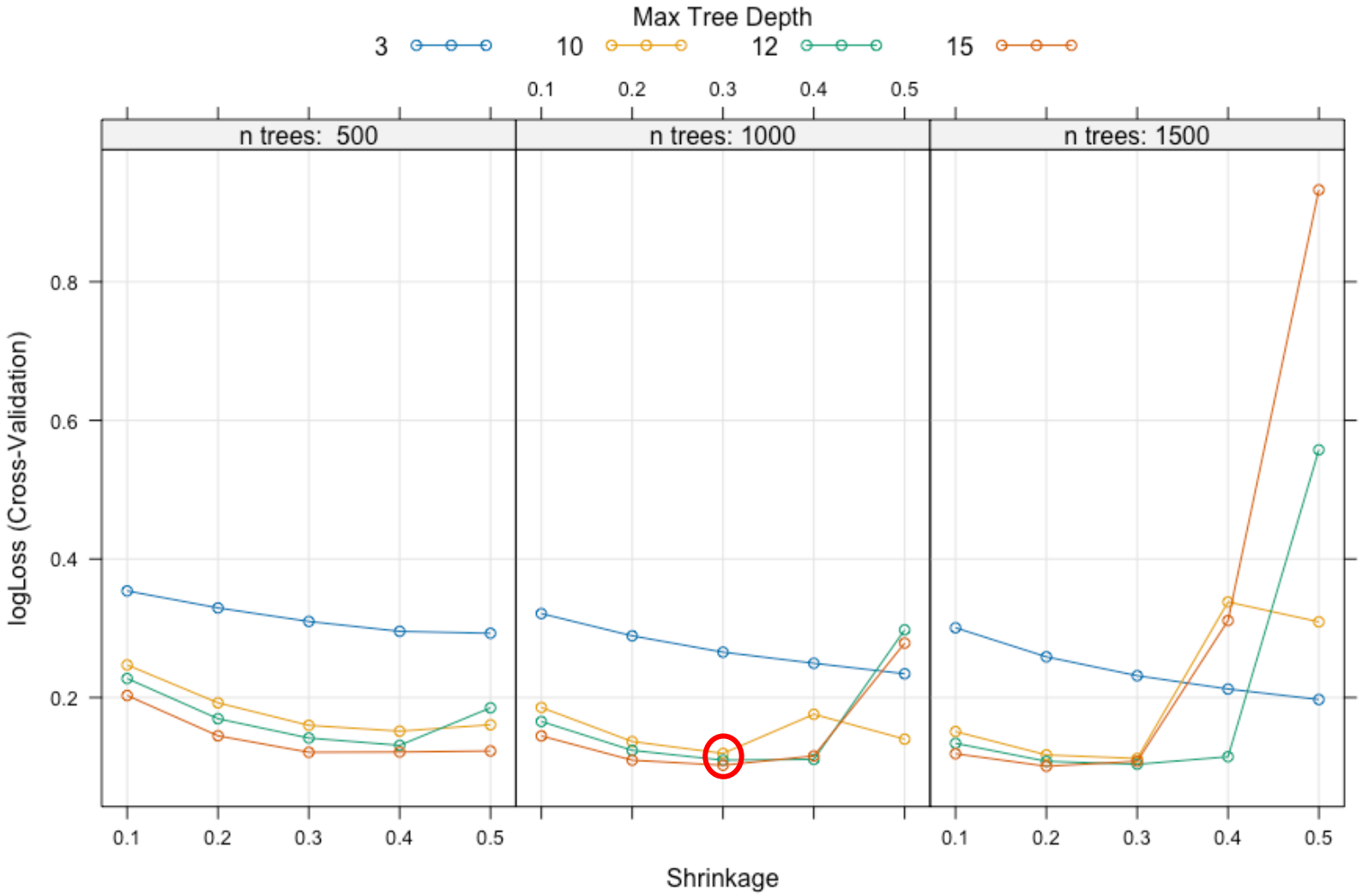
Model (4/4) : Boosting Trees

$\max tree\ depth$ = 15

$shrinkage$ = 0.3

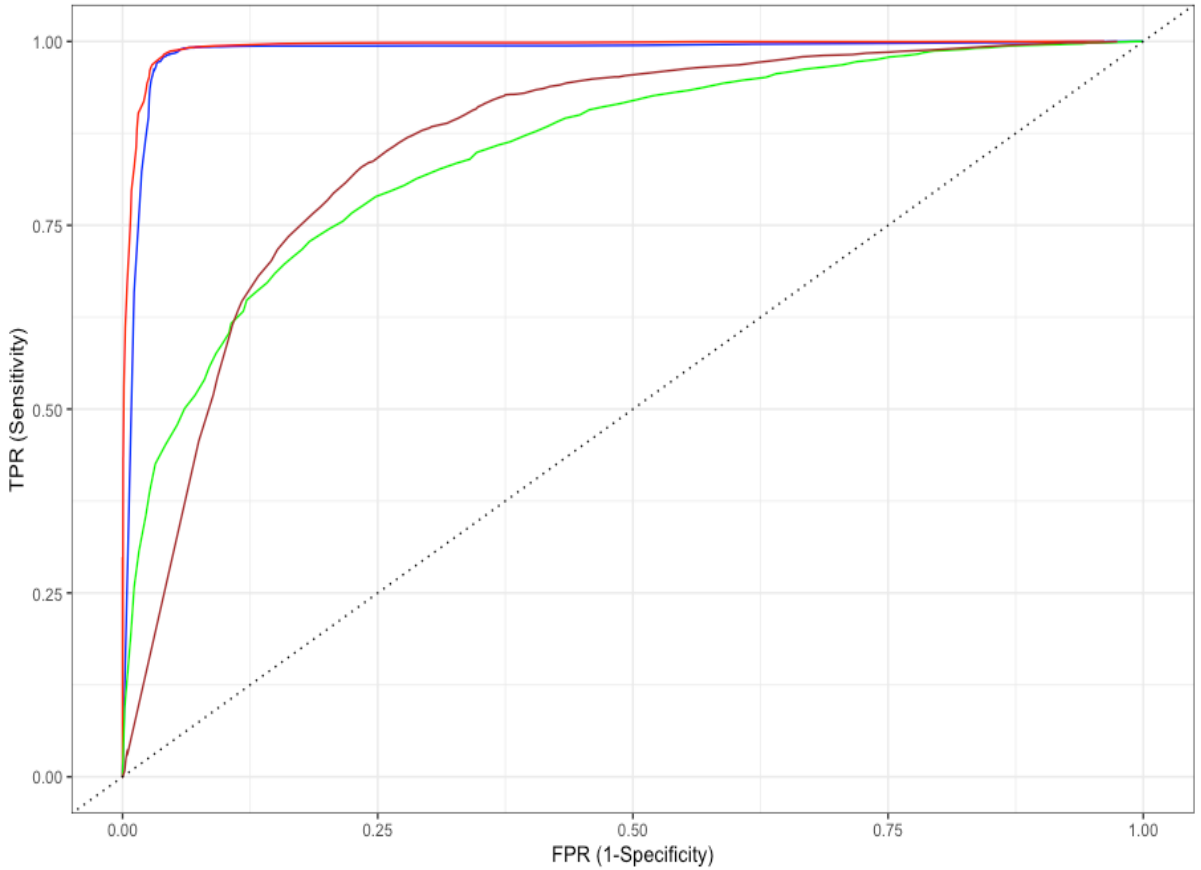
$\#\ Trees$ = 1000

	Reference	
Prediction	No	Yes
No	72.5	1.5
Yes	1.0	25.1

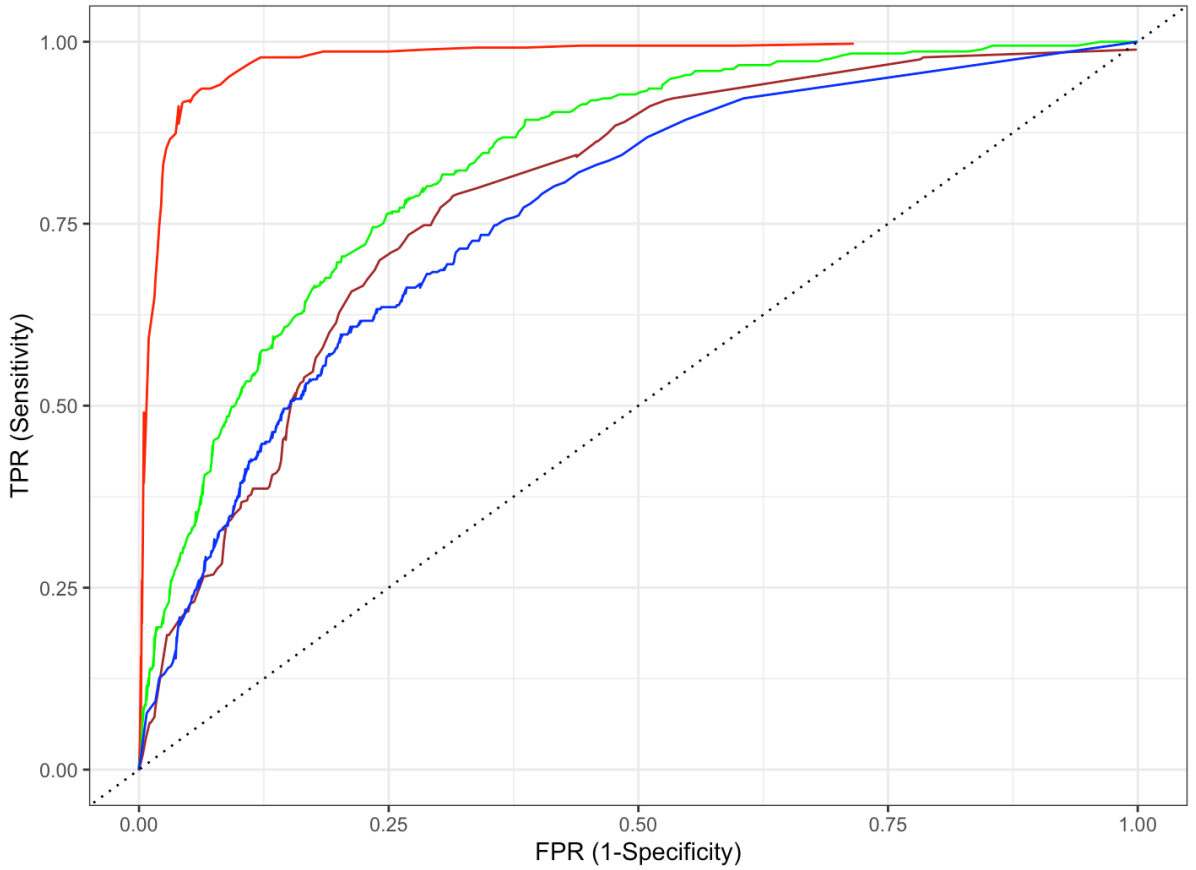


Performance Comparison

Cross Validation sets



Holdout



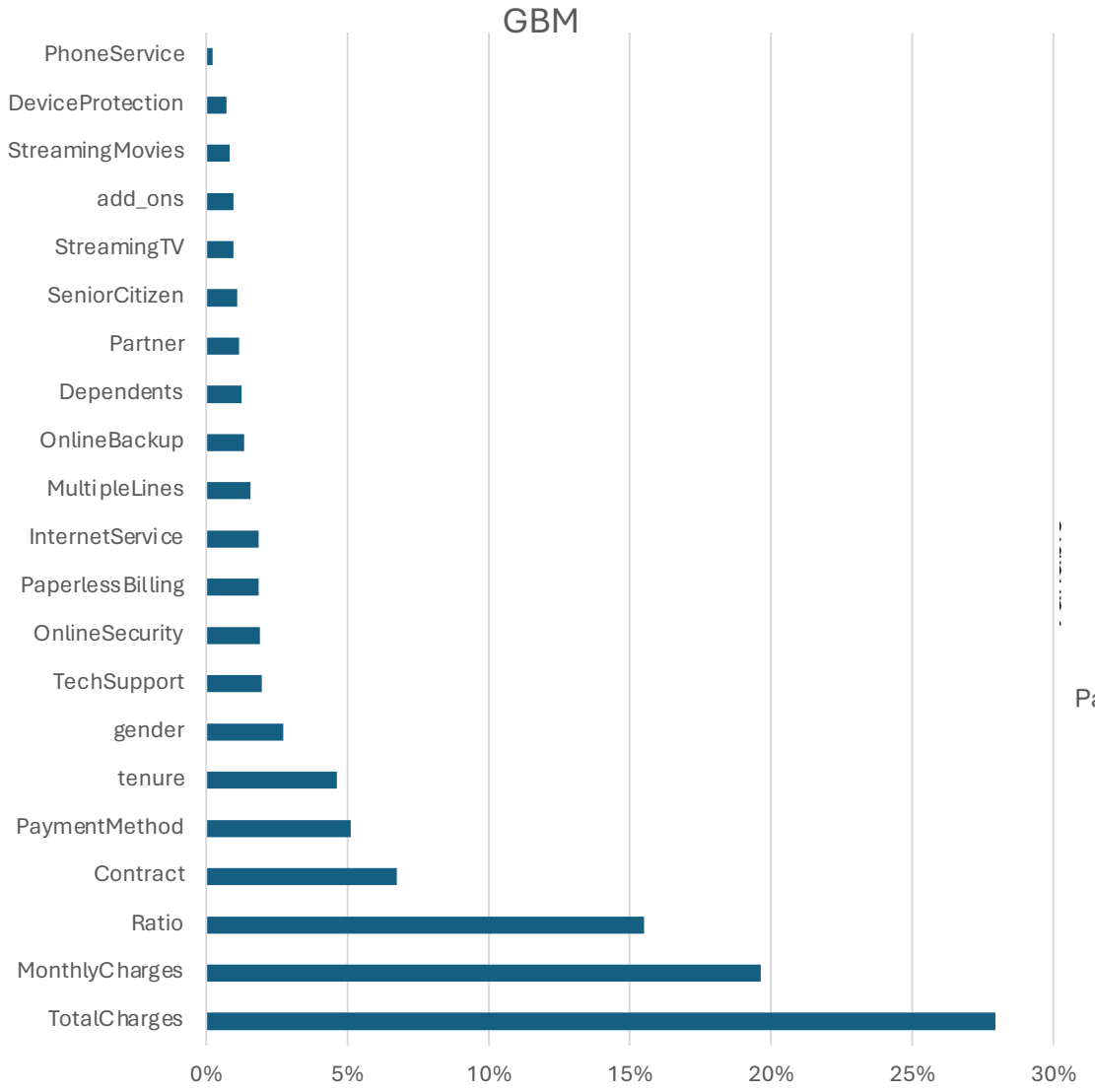
Random Forest

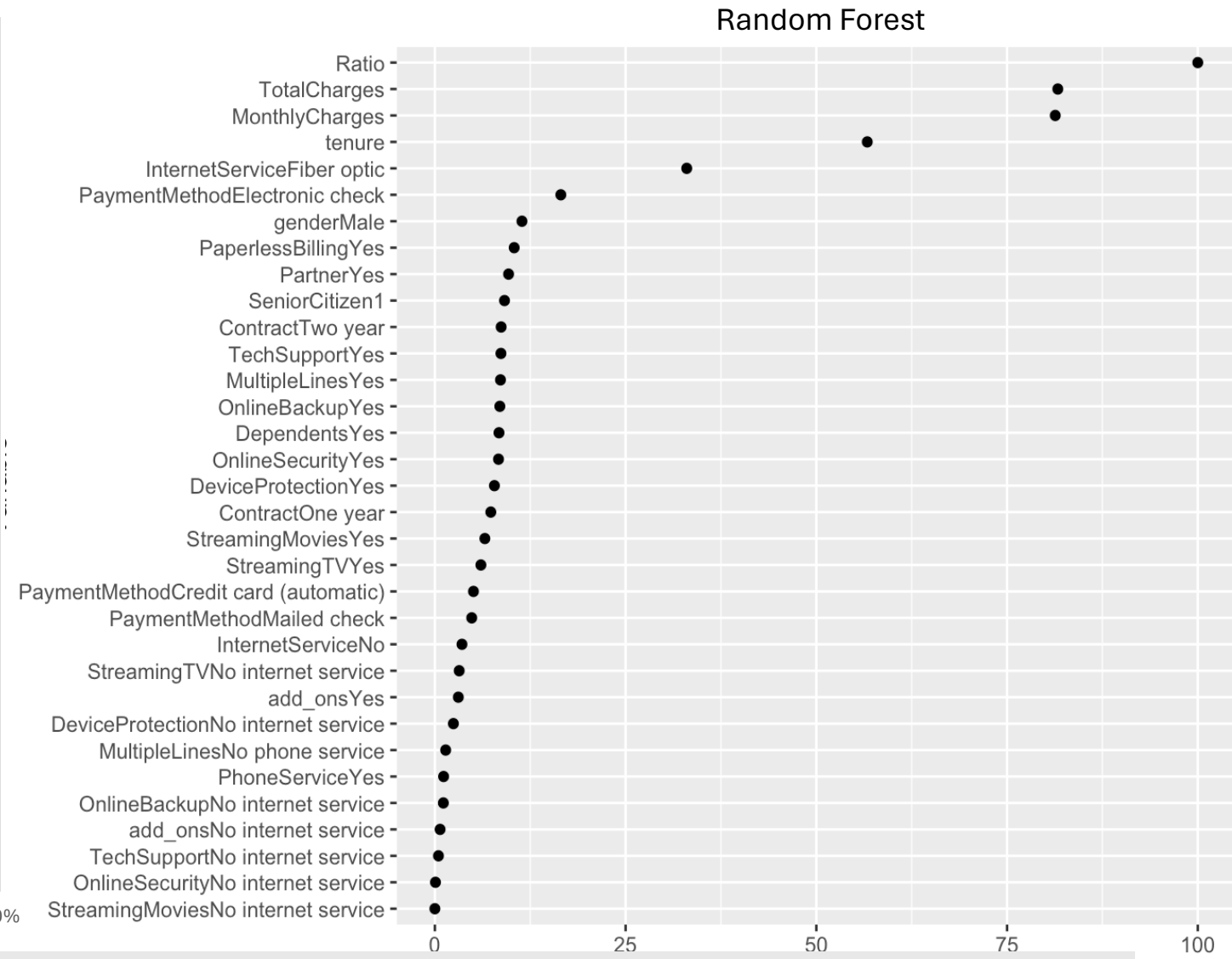
GBM

Logistic Regression

Single Decision Tree

Variable Importance



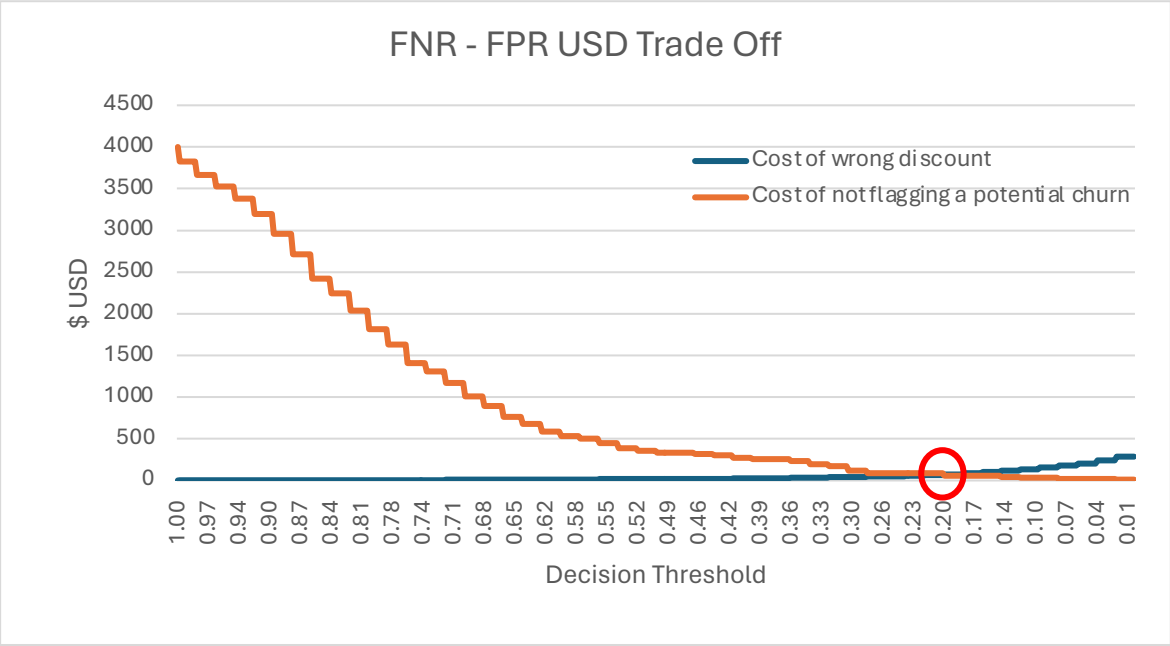


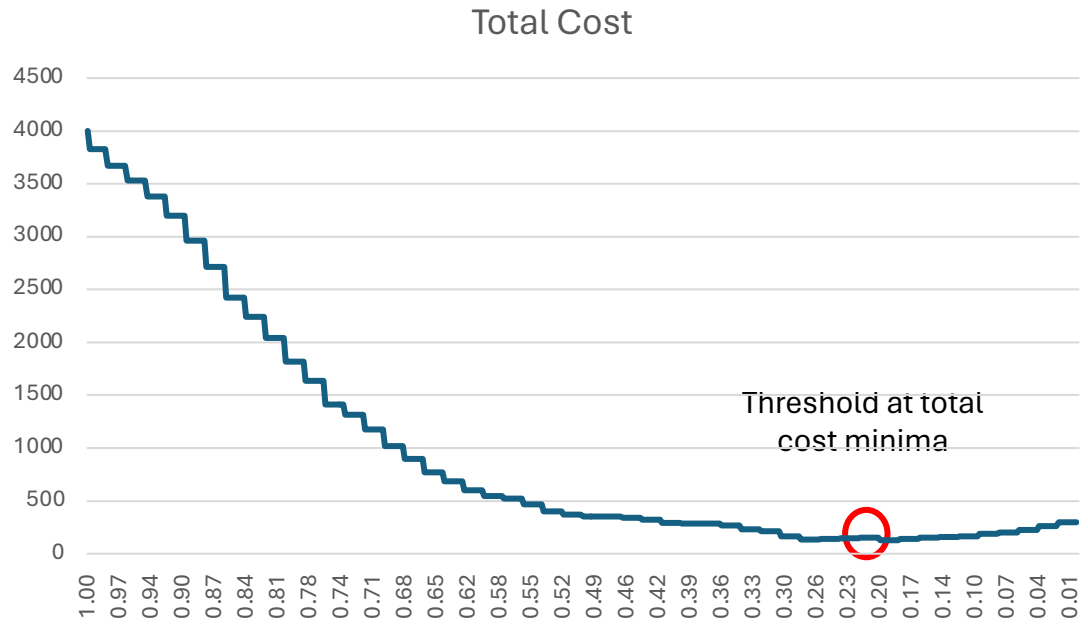
Which variables are important across both ?

Threshold Methodology

- Finding the right threshold is a trade off between
  - Cost of giving discounts to the incorrect customer (FPR) v / s
  - Customer lifetime value loss incurred by not identifying a true churn (FNR)

$$Total\ Cost = FPR\ *(\$\ Retention\ Discount) + FNR\ *(\$\ Lost\ customer\ LTV)$$





- Assumption – Retention cost / customer is 10% of Customer LTV

Reference Excel Sheet

Threshold Methodology

Performance on Holdout Data

Decision Threshold – 0.2

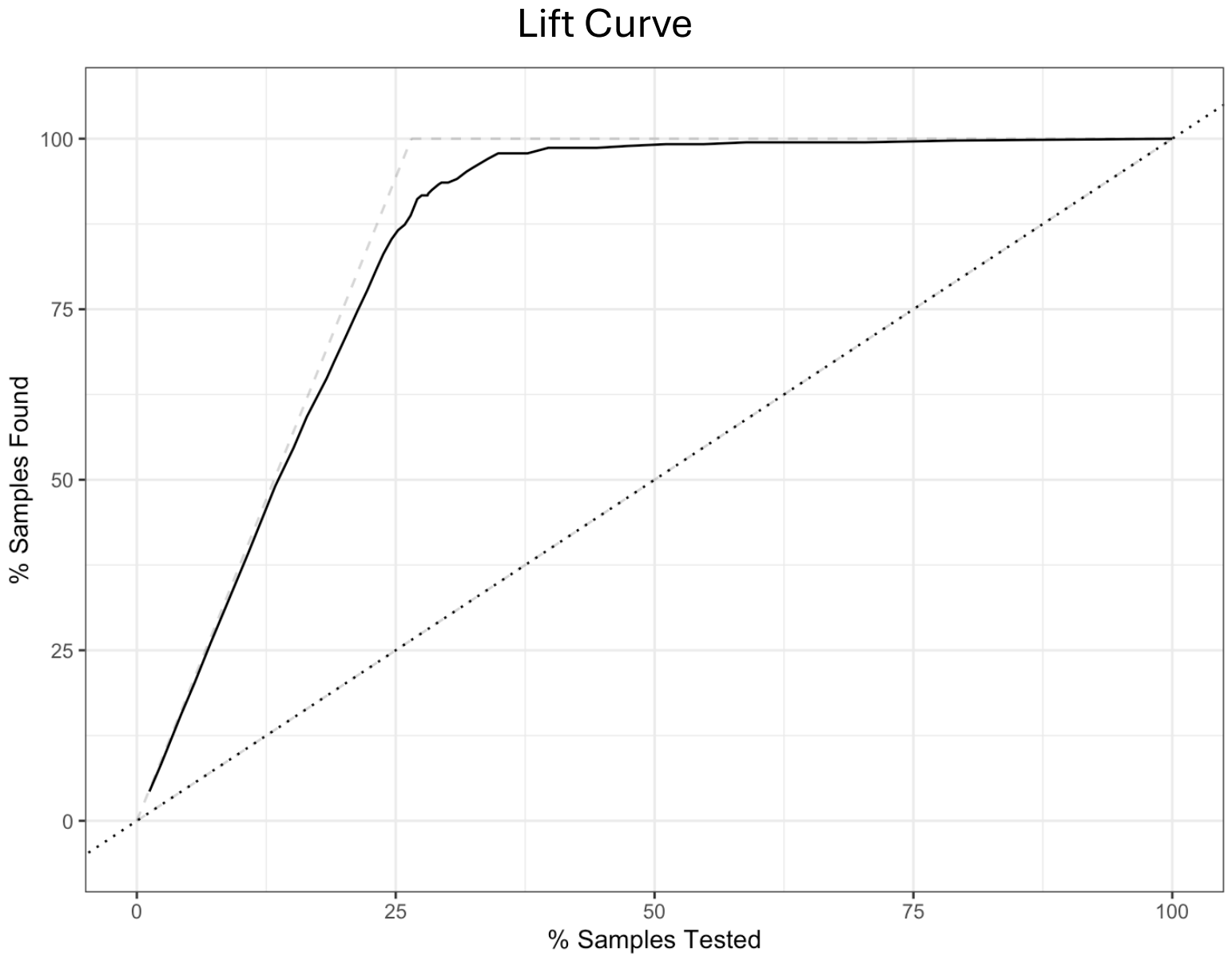
```
Confusion Matrix and Statistics

          Reference
Prediction Yes  No
       Yes 326 357
       No   47 677
```

Precision	47%
Recall (Capture Rate)	87%

- Cost of false negative (missing out a future churn) is **much higher** than the cost of a false positive (giving discount to the wrong customer)
- Business requirements dictate higher preference to recall over precision

Business Actionable Insights



Using the model,

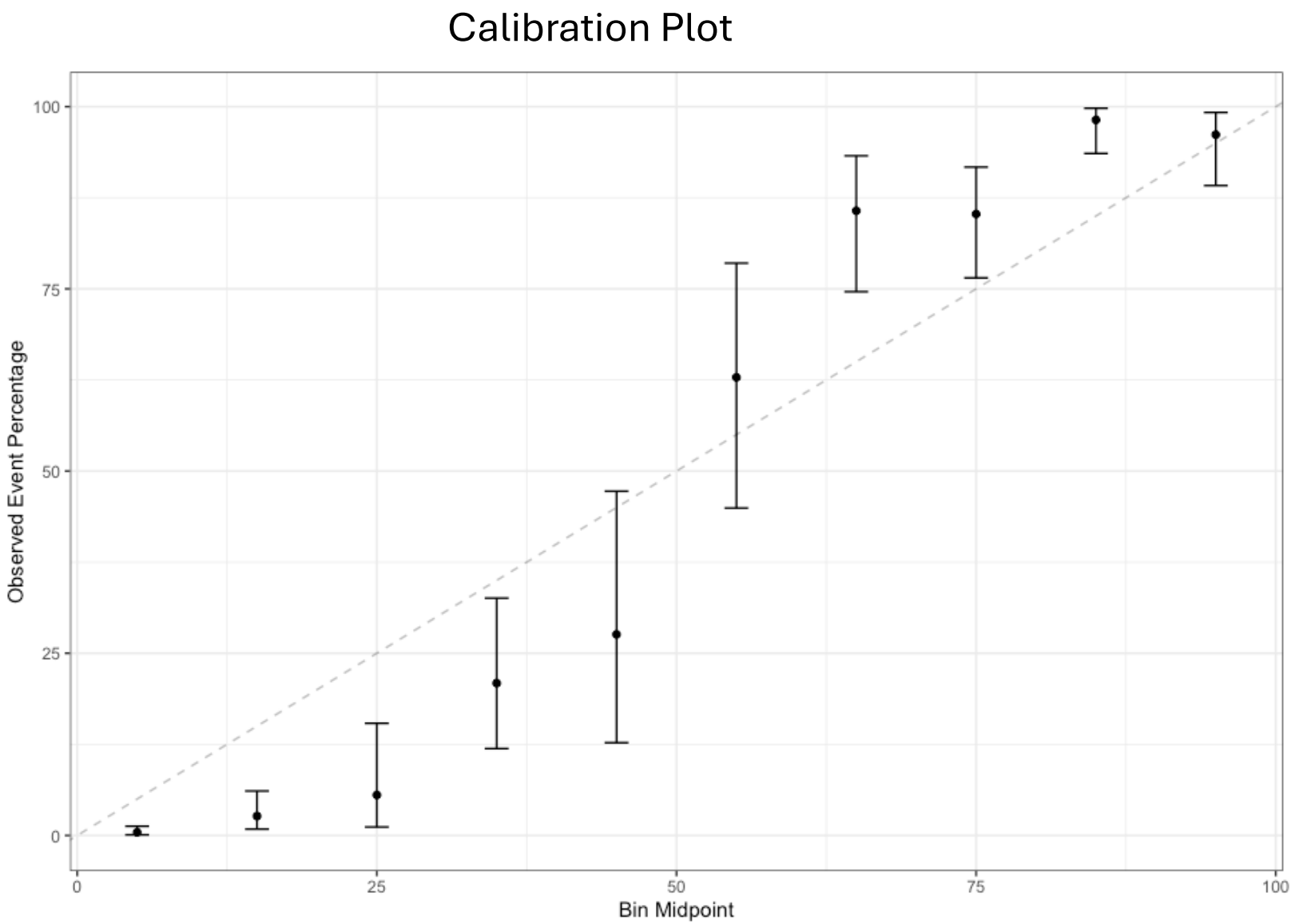
80%

of the potential churns can be saved using

only 25%

of the retention budget

Appendix : Scope for Improvement



- The model slightly over predicts in the lower deciles of scores
- And under predicts in the higher ones
- Scope for re-calibration using advanced techniques

References

BlastChar. (2018, February 23). *Telco customer churn*. Kaggle. https://www.kaggle.com/datasets/blastchar/telco-customer-churn

Sabrina TessitoreSabrina is Content Marketing Manager and qualified B2B AX-pert at CustomerGauge. She provides the strategies necessary for B2B companies to build ROI-generating NPS programs. In Sabrina's free time, & Tessitore, S. (n.d.). *What's the average churn rate by industry?* CustomerGauge. https://customergauge.com/blog/average-churn-rate-by-industry

Virgin Media O2 sets up for 2024 execution with focused investments in Q1. (n.d.). https://news.virginmediao2.co.uk/wp-content/uploads/2024/05/Virgin-Media-O2-Q1-2024-Earnings-Release.pdf