# Predicting Customer Churn

Noah Armsworthy
November 8th, 2024

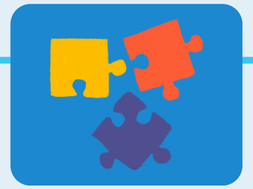
# Problem Statement

To maintain the capital required to provide security to their customers, insurance companies must retain as many customers as possible. Predicting potential customer churn can help inform companies of critical points where retention strategies can be deployed.

# METHODOLOGY

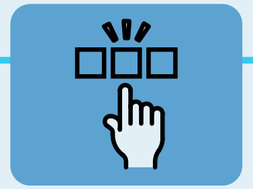
## INITIAL ANALYSIS

- Getting to know the data



## DATA PREPARATION

- Improve data quality
- Remove potentially misleading patterns



## MODEL SELECTION

- Metric comparison on default parameters
- Cross validation

## MODEL TRAINING & TUNING

- Hyper parameter tuning
- Interpret results

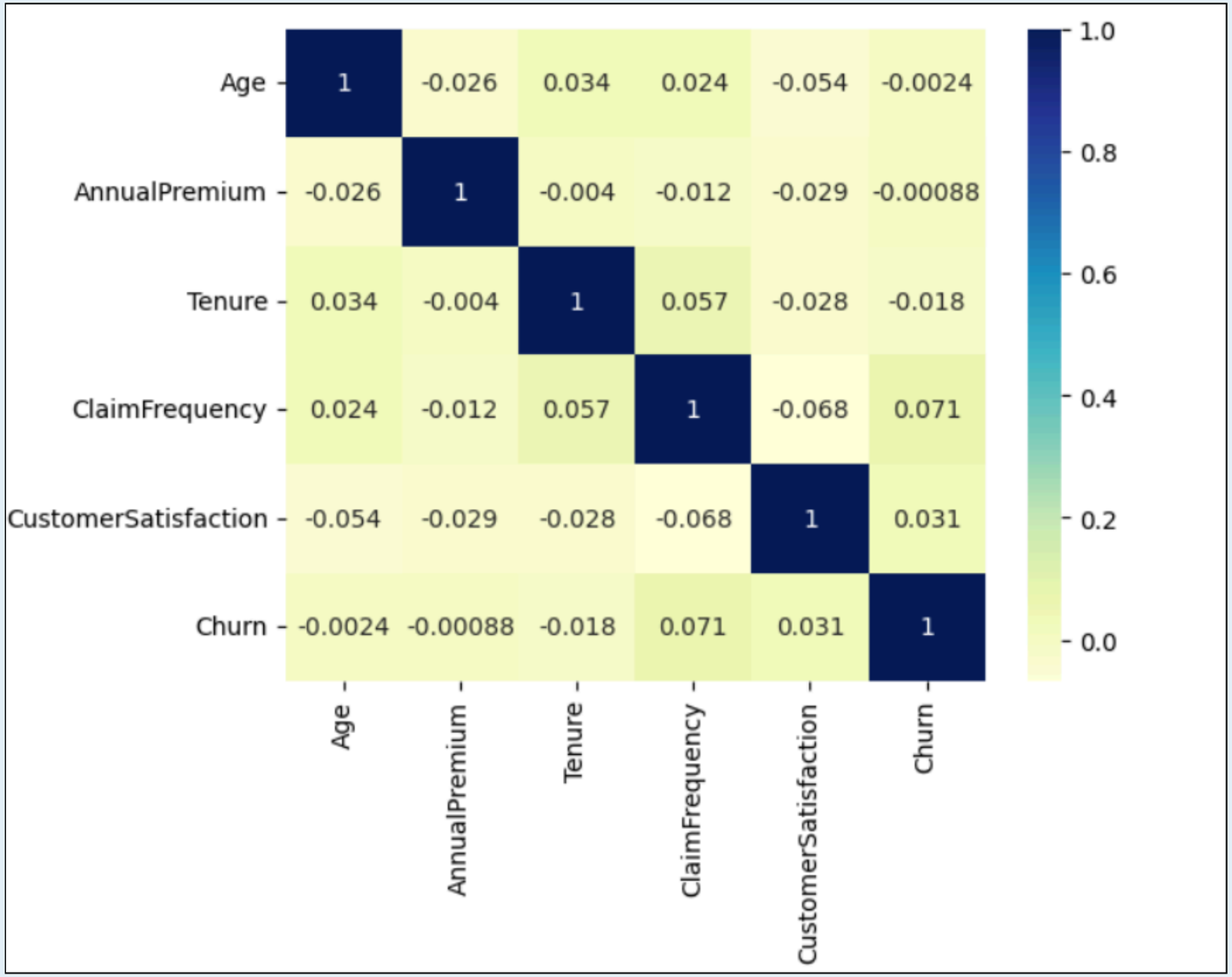
# Initial Analysis

| Feature Name | Distinct | Missing | Min | Max | Mean | Std |
|---|---|---|---|---|---|---|
| Customer ID | 1193 (20%) | 0 | 1001 | 2200 | N/A | N/A |
| Age | 81 (1%) | 593 (10%) | 13 | 125 | 50 | 18 |
| Annual Premium | 1241 (21%) | 591 (10%) | 0 | 9693 | 1175 | 630 |
| Tenure | 29 (21%) | 600 (10%) | 1 | 29 | 15 | 8.4 |
| Province | 5 (<1%) | 0 | N/A | N/A | N/A | N/A |
| Marital Status | 3 (<1%) | 0 | N/A | N/A | N/A | N/A |
| Policy Type | 3 (<1%) | 0 | N/A | N/A | N/A | N/A |
| Claim Frequency | 1193 (20%) | 0 | 0 | 4.9 | 2.5 | 1.4 |
| Customer Satisfaction | 10 (<1%) | 0 | 1 | 10 | 5.6 | 2.8 |

# Initial Analysis

## Correlation

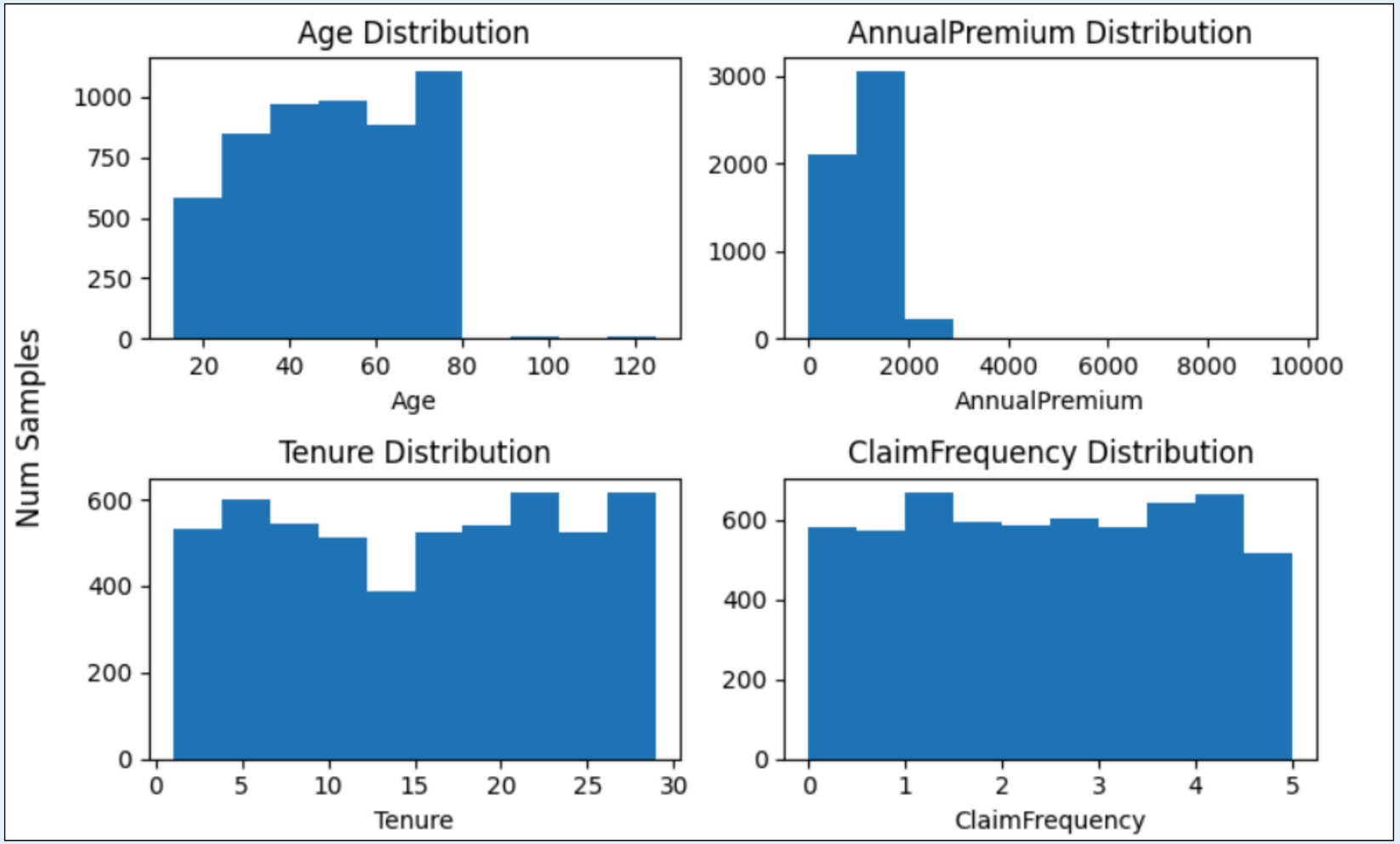
| | |
|---|---|
| **0.0 to +0.2** | **Very weak + or no association** |
| 0.0 to -0.2 | Very weak - or no association |




https://sphweb.bumc.bu.edu/otlt/MPH-Modules/PH717-QuantCore/PH717-Module9-Correlation-Regression/PH717-Module9-Correlation-Regression4.html

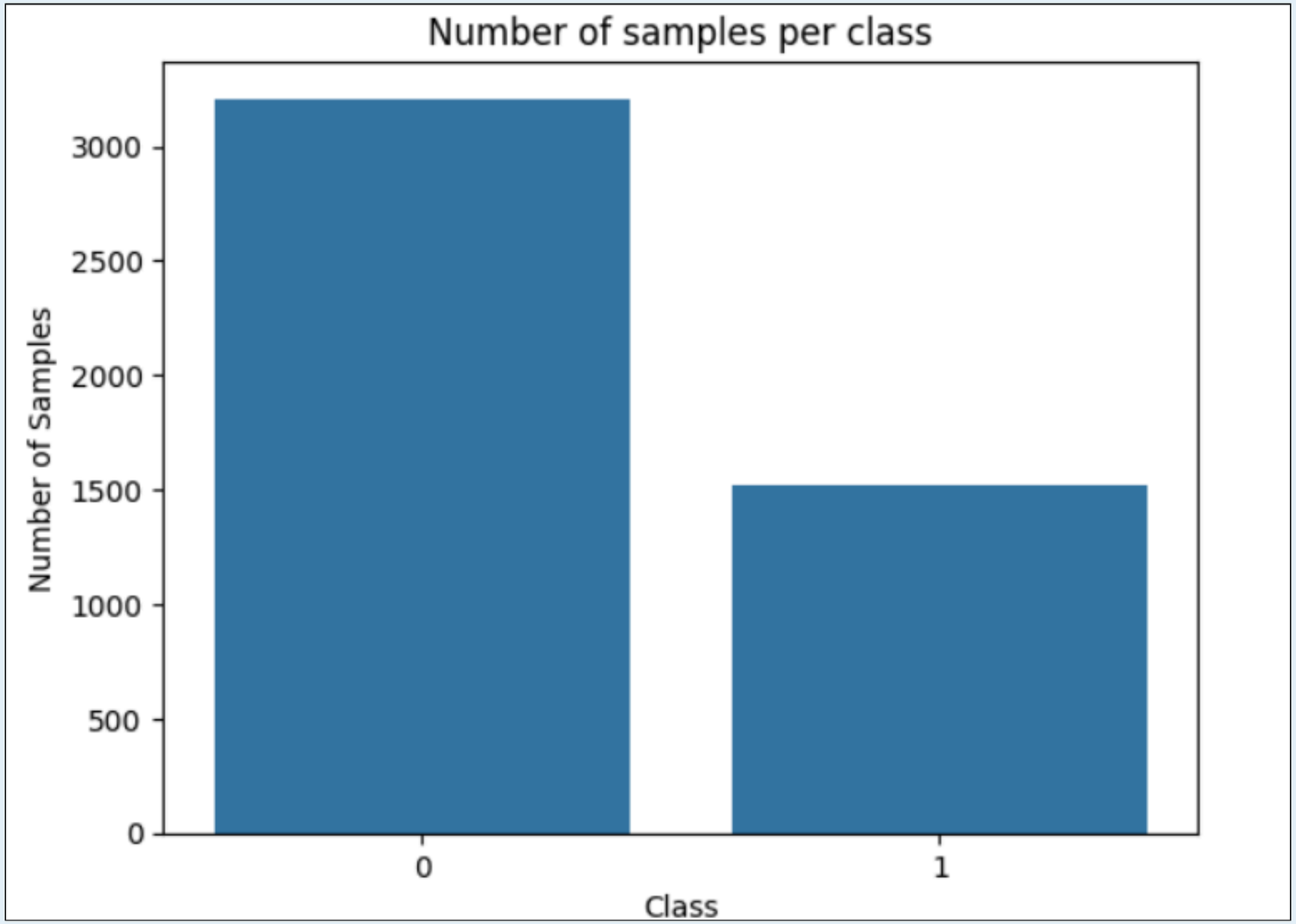
# Initial Analysis

## Distributions



# Initial Analysis

## ~2:1 Imbalance for No Churn



# Data Preparation

Missing Data

- **Less than 0.01 correlation to the target column**
- **5% Rule of Thumb**
- **Use simple imputer to place the most common value in that column**

# Data Preparation

Duplicate Records

- **Near duplicates provide conflicting information for the same sample**
- **If valid, entries need a date or ‘active’ column**
- **Ultimately, dropping these would lose too much data.**

**Exact duplicates were dropped.**
**Near duplicates were kept.**

# Data Preparation

## Min-Max Scaling

| | Age | AnnualPremium | Tenure | ClaimFrequency | CustomerSatisfaction |
|---|---|---|---|---|---|
| **count** | 4721.000 | 4721.000 | 4721.000 | 4721.000 | 4721.000 |
| **mean** | 0.427 | 0.108 | 0.563 | 0.499 | 0.512 |
| **std** | 0.236 | 0.076 | 0.318 | 0.286 | 0.315 |
| **min** | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 |
| **25%** | 0.229 | 0.056 | 0.286 | 0.251 | 0.222 |
| **50%** | 0.434 | 0.108 | 0.571 | 0.500 | 0.556 |
| **75%** | 0.651 | 0.157 | 0.857 | 0.750 | 0.778 |
| **max** | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 |

| Model | Accuracy | Precision | | Recall | | F1-Score | | ROC AUC |
|---|---|---|---|---|---|---|---|---|
| | | No Churn | Churn | No Churn | Churn | No Churn | Churn | |
| Logistic Regression | 67% | 67% | 0% | 100% | 0% | 80% | 0% | 55% |
| SVM | 67% | 67% | 0% | 100% | 0% | 80% | 0% | 55% |
| Decision Tree Classifier | 90% | 91% | 87% | 94% | 82% | 93% | 84% | 87% |
| K-Nearest Neighbor | 64% | 69% | 43% | 84% | 25% | 76% | 31% | 57% |
| Naive Bayes | 66% | 67% | 40% | 98% | 3% | 80% | 5% | 53% |
| Random Forest Classifier | 93% | 90% | 100% | 100% | 78% | 95% | 88% | 98% |
| Gradient Boosting Classifier | 73% | 71% | 87% | 98% | 21% | 83% | 34% | 84% |

# Model Selection

# Model Selection

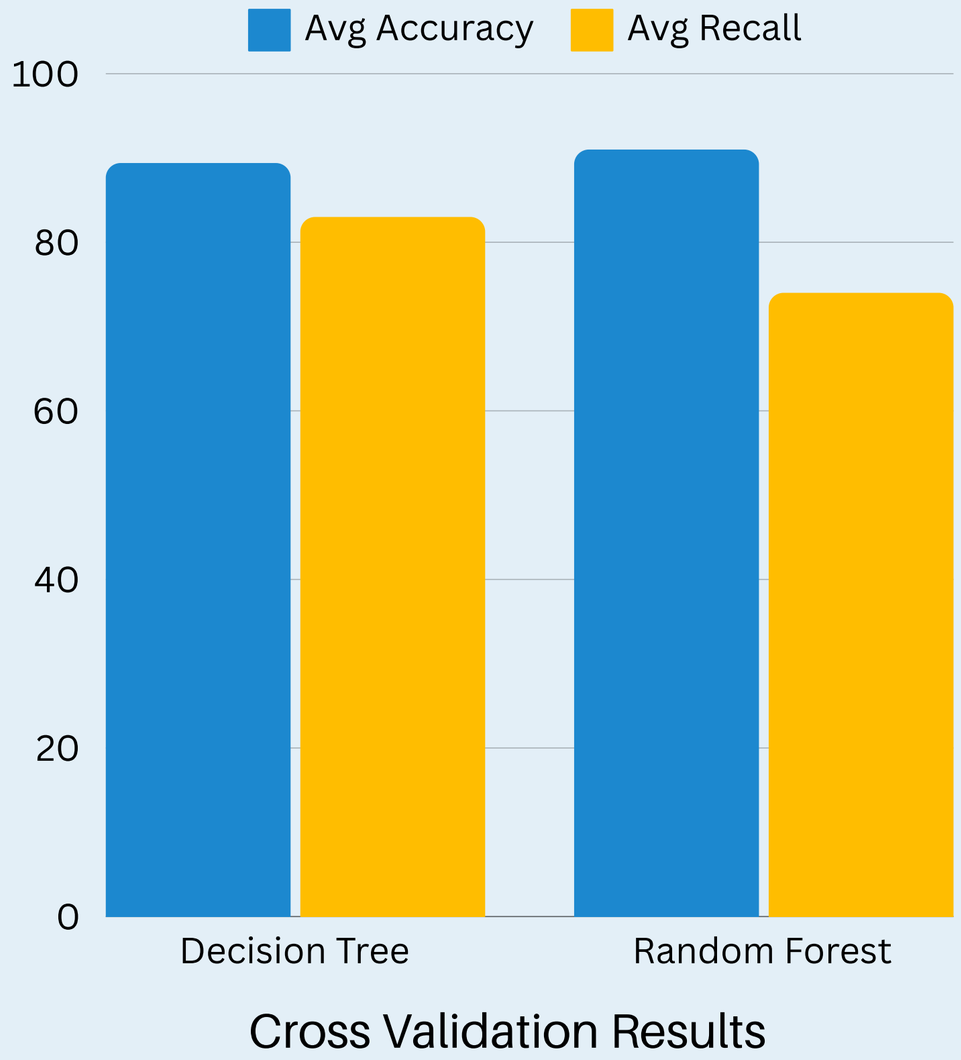
## Decision Tree

$$Accuracy = \frac{CorrectPredictions}{AllLabels}$$

$$Recall = \frac{CorrectPositivePredictions}{AllPositiveLabels}$$

Additional benefits of Decision Tree:

- Faster training and tuning
- Potentially more interpretable
- Good for small datasets



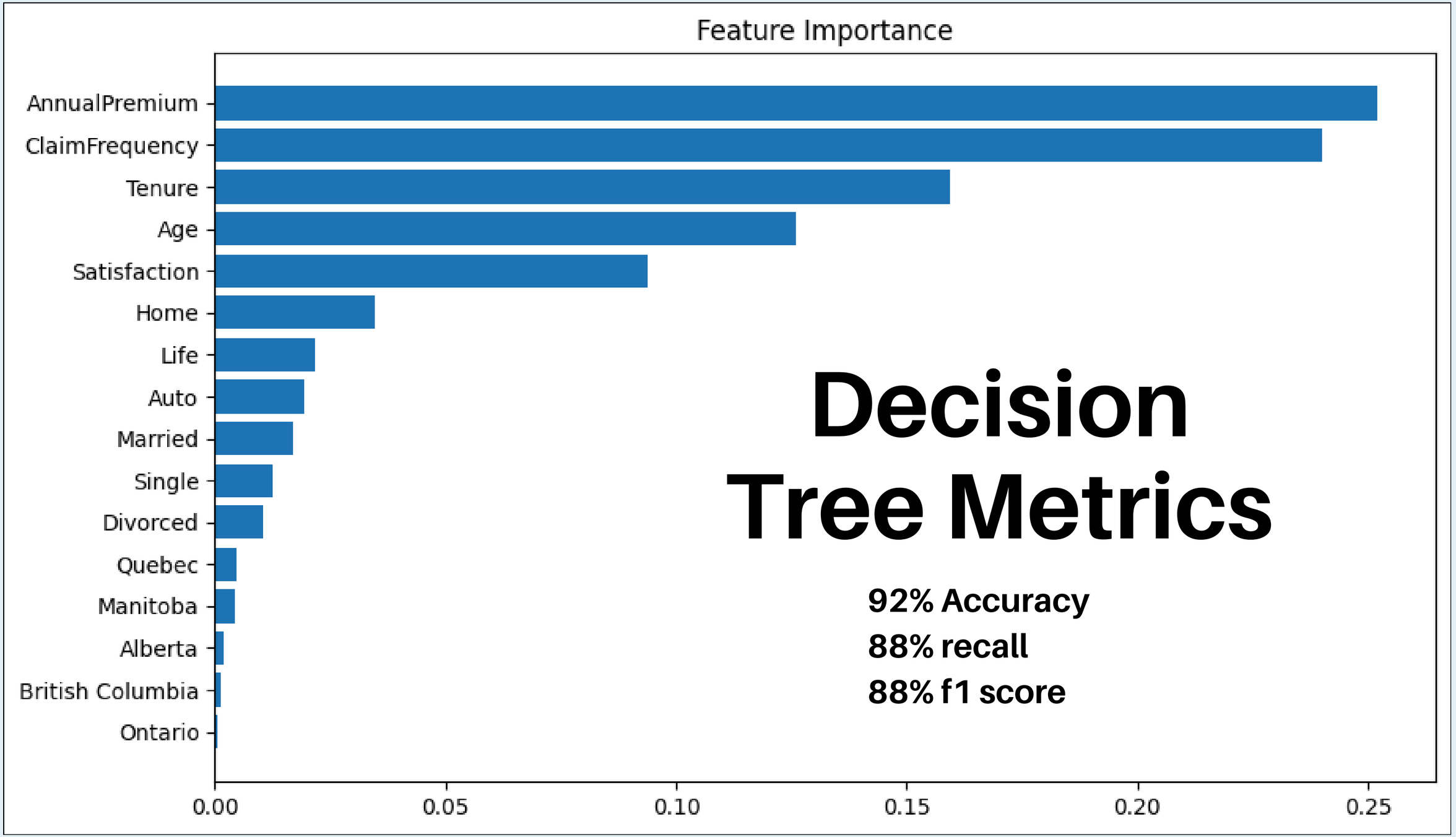

Cross Validation Results

# Decision Tree Training & Tuning

Chose **f1** scoring as the *harmonic mean* between precision and recall.
“How **good** and **complete** are the positive predictions?”

## Using Grid Search Cross Validation

- Class weight: {No Churn:1, Churn: **1.5**}
- Criterion: ‘**gini**’
- **No** specific max depth or max features.
- Minimum samples per leaf: **1**
- Minimum required samples to split: **2**
- Average **f1** score of **89%**


Feature Importance
AnnualPremium
ClaimFrequency
Tenure
Age
Satisfaction
Home
Life
Auto
Married
Single
Divorced
Quebec
Manitoba
Alberta
British Columbia
Ontario
0.00
0.05
0.10
0.15
0.20
0.25
Decision Tree Metrics
92% Accuracy
88% recall
88% f1 score

# Challenges & Next Steps

**01** Consult with the data provider about duplicates records and possible outliers.

**02** Consider over/under-sampling if future data has different class distributions.

**03** Gather more relevant data.

**04** Post-pruning for the Decision Tree may improve performance and interpretability.

**05** Perform testing on a larger variety of hyper parameters.


# Thank you!

**Noah Armsworthy**

noaharmsworthy@gmail.com