Predicting **Churn** with **Machine Learning** Algorithms

Group 13

Akila Herath[1]
✉ akila.herath@postgrad.plymouth.ac.uk

Subina Maharjan[1]
✉ subina.maharjan@postgrad.plymouth.ac.uk

Ei Ei Maw[1]
✉ ei.maw@postgrad.plymouth.ac.uk

Emad Farjami[1]
✉ emad.farjami@postgrad.plymouth.ac.uk

[1] *School of Engineering, Computing and Mathematics (Faculty of Science and Engineering) | University of Plymouth, UK*

Introduction

The churn rate, a key business metric, indicates customer loyalty and engagement **(Investopedia 2023)**. It is calculated by dividing the number of churned customers by the total customers at the start of the period **(Luck 2023)**. High churn rates, particularly in telecommunications and subscription services, can significantly impact profitability and growth. Each industry has unique challenges and retention strategies, as shown in the industry-wide churn rates in figure 1.

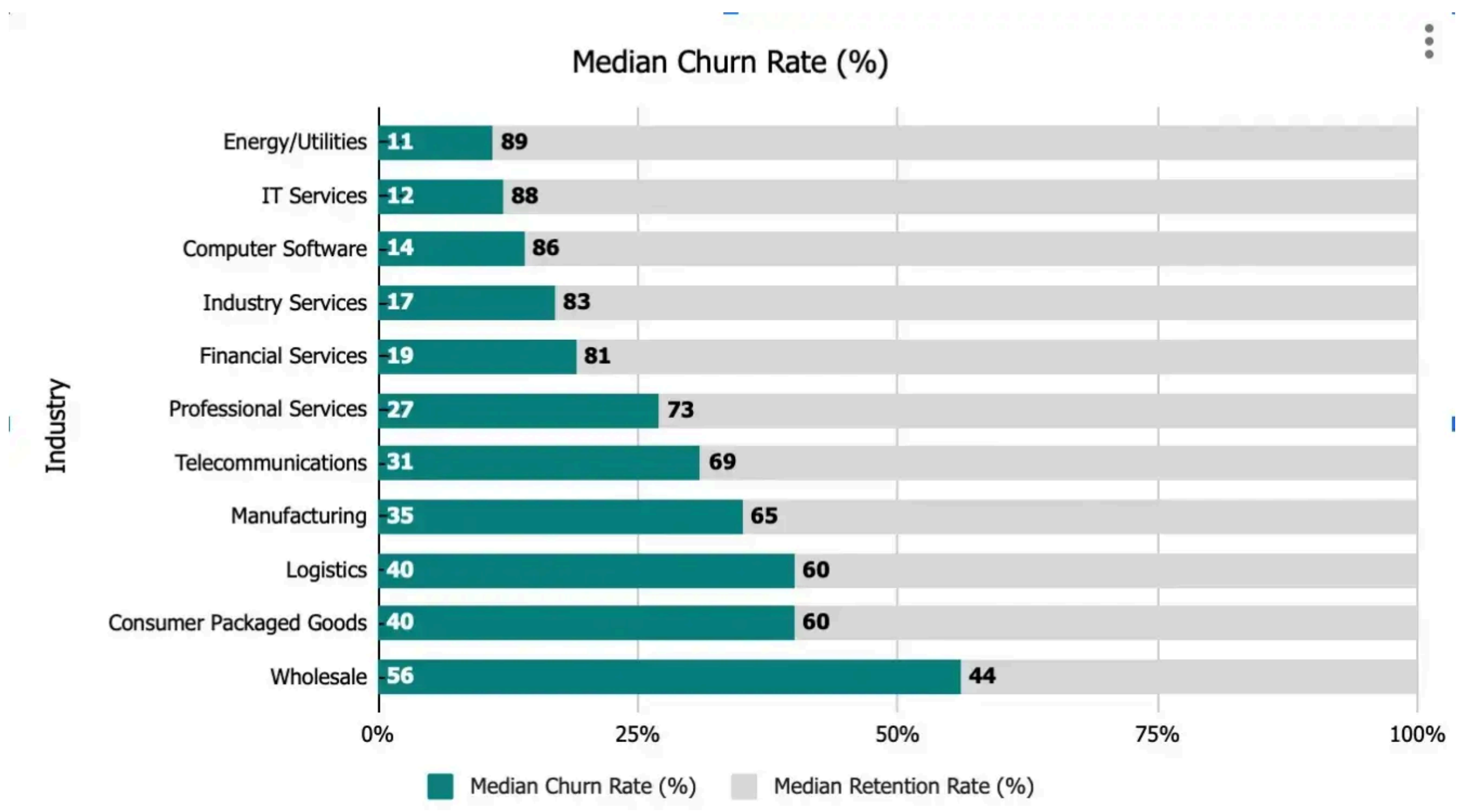


Figure 1: Median Customer Churn Rate by Industry 2022 **(Tessitore 2023)**

Not managing churn can lead to a 25% loss of current customers, translating into an annual revenue loss of 75 billion **(Shabankareh et al. 2021)**. Industries like insurance, telecommunications, and credit cards, facing intense competition, are particularly prone to churn, where even a 1% reduction in churn can increase profits by 6% **(Shabankareh et al. 2021)**.

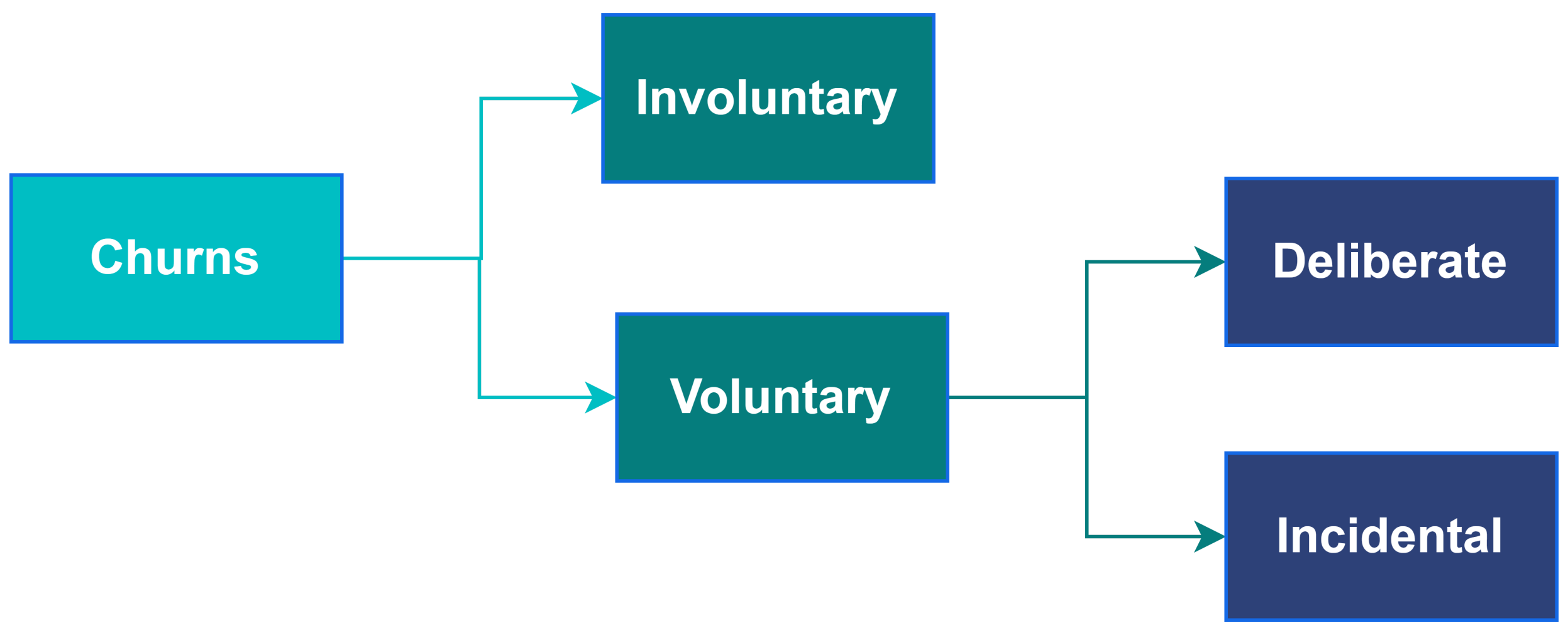

Churns
Involuntary
Voluntary
Deliberate
Incidental

Figure 2: Customer Churn Prediction Analysis **(Ly and Son 2022)**

Previous Studies on Churn Prediction

(#tab:table 0)Summary of Studies

Author	Dataset	Methods	Accuracy.Rate
Ullah et al.	South Asia GDM telecom (CDR) 64k instances, 29 attrs, 17 features	Random Forest	88.63%
Beeharry & Fokone	IBM Sample & Duke University, 21 & 57 features	Ensemble (KNN, RF, LR, NB)	82.30%, 63% F1 (imbal.), 76.20%, 77.06% F1 (bal.)
Bilisik & Sarp	IBM's Open access, 21 features	ANN, SVM, RF	82% (ANN), 79% (SVM), 80% (RF)
Ahmad et al	Syrian Telecom (SyriaTel), 2000 features	DT, RF, GBM, XGBoost	93.3% (XGBoost, unbal.)

Our Dataset

The dataset, sourced from Kaggle, comprises 7043 subscribers. Each row corresponds to a customer, and each column denotes their characteristics, as depicted in Figure 3.

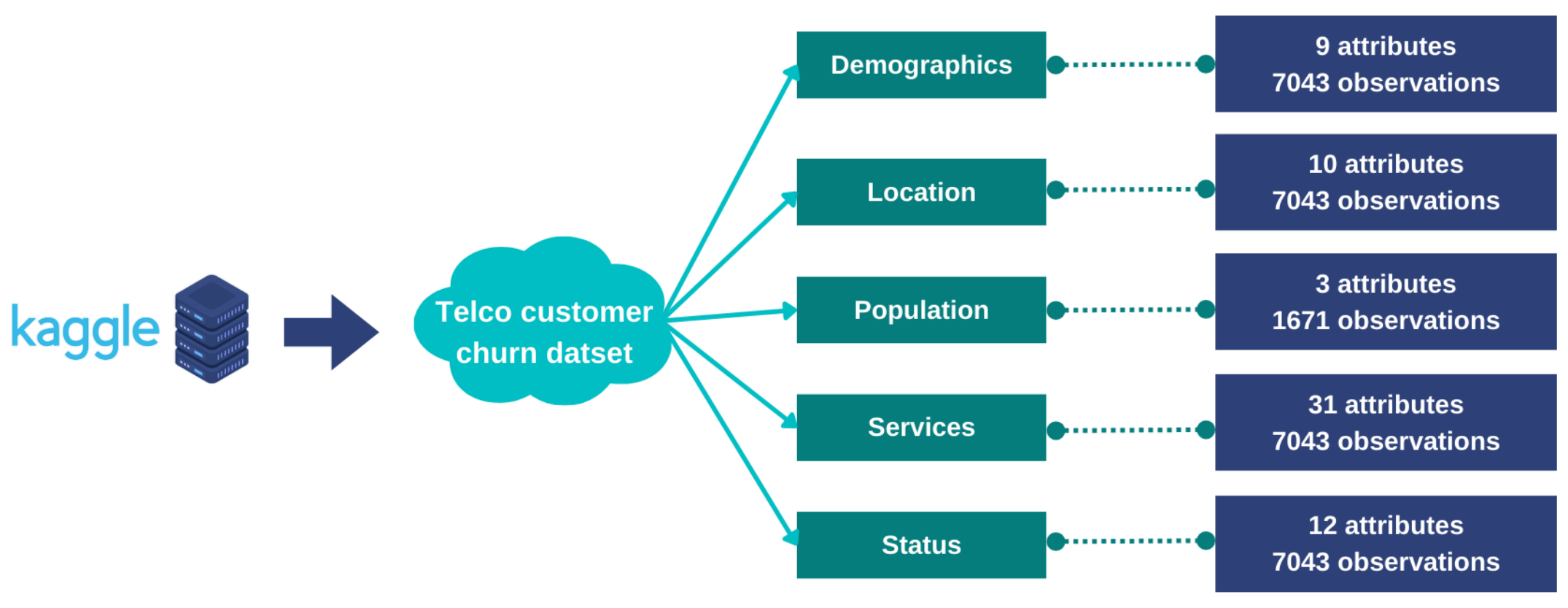

kaggle
Telco customer churn datset
Demographics — 9 attributes, 7043 observations
Location — 10 attributes, 7043 observations
Population — 3 attributes, 1671 observations
Services — 31 attributes, 7043 observations
Status — 12 attributes, 7043 observations

Figure 3: Kaggle Dataset **(Bansal 2023)**

Our Machine Learning Model

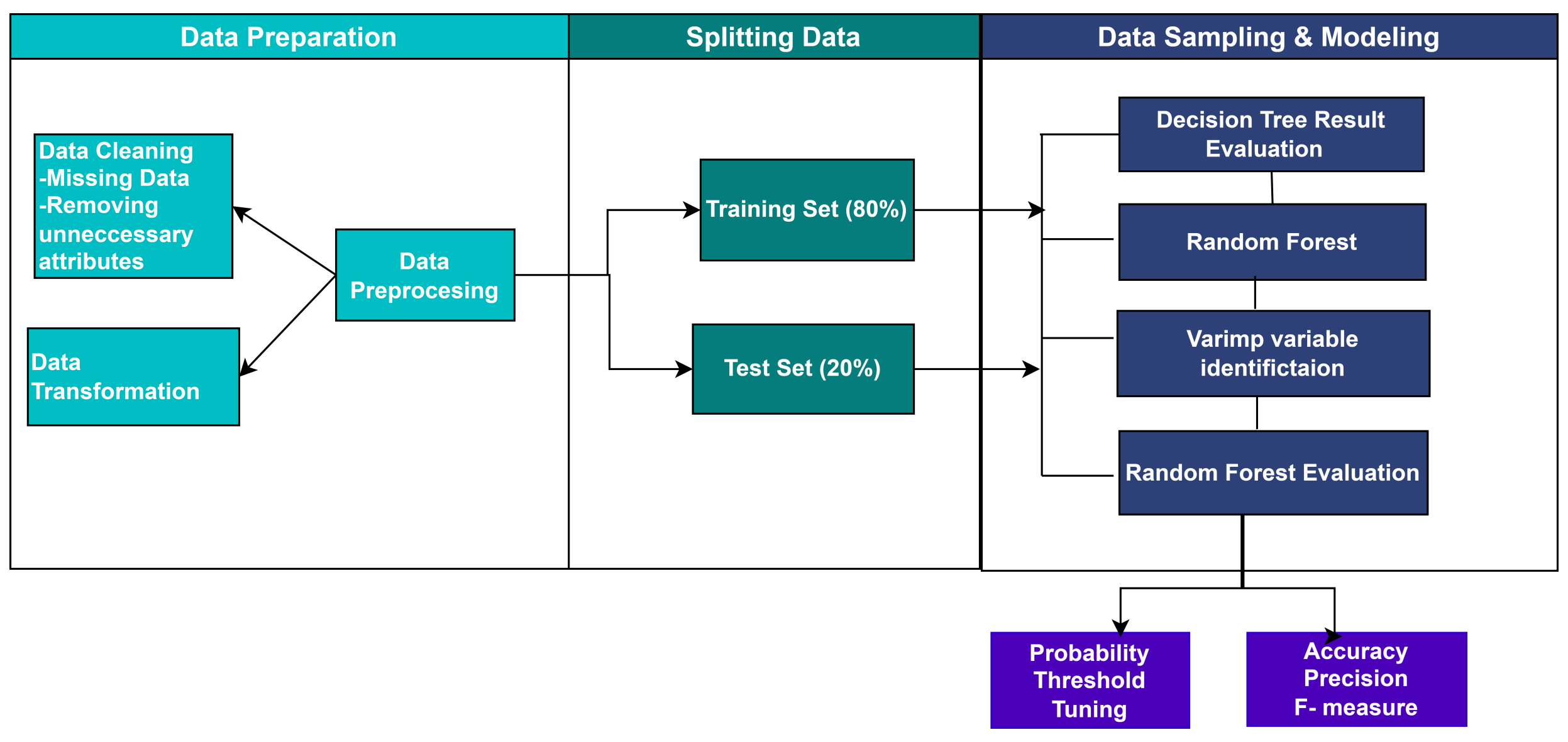

Data Preparation
Splitting Data
Data Sampling & Modeling
Data Cleaning
-Missing Data
-Removing unneccessary attributes
Data Transformation
Data Preprocesing
Training Set (80%)
Test Set (20%)
Decision Tree Result Evaluation
Random Forest
Varimp variable identifictaion
Random Forest Evaluation
Probability Threshold Tuning
Accuracy Precision F- measure

Figure 4: ML Model

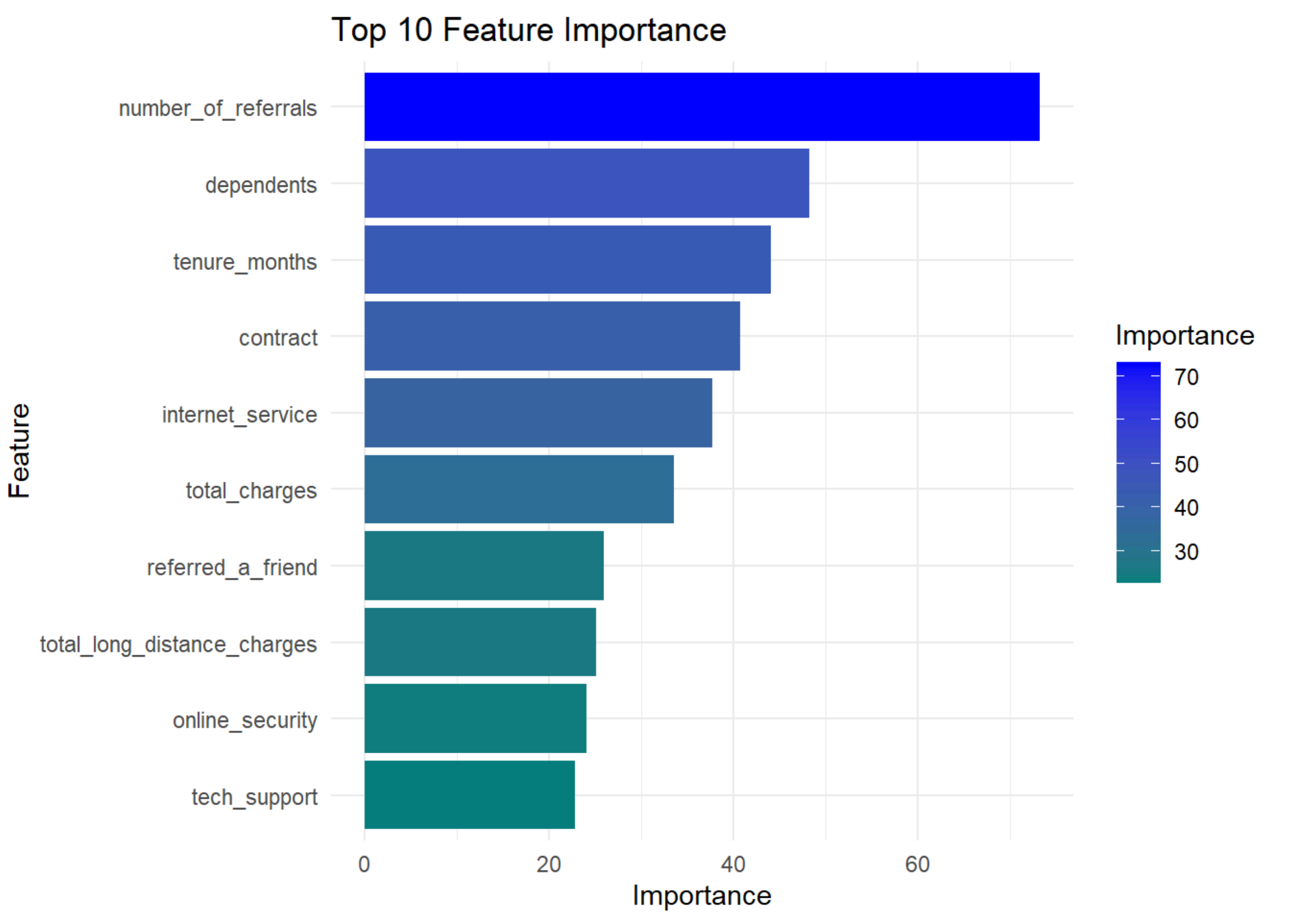
(Bilisik and Sarp 2023) utilized the same dataset for their investigation, where the class distribution wasn't uniform. They conducted model runs with three distinct training sets employing oversampling and undersampling techniques. The outcomes of this analysis are detailed in Table 2.

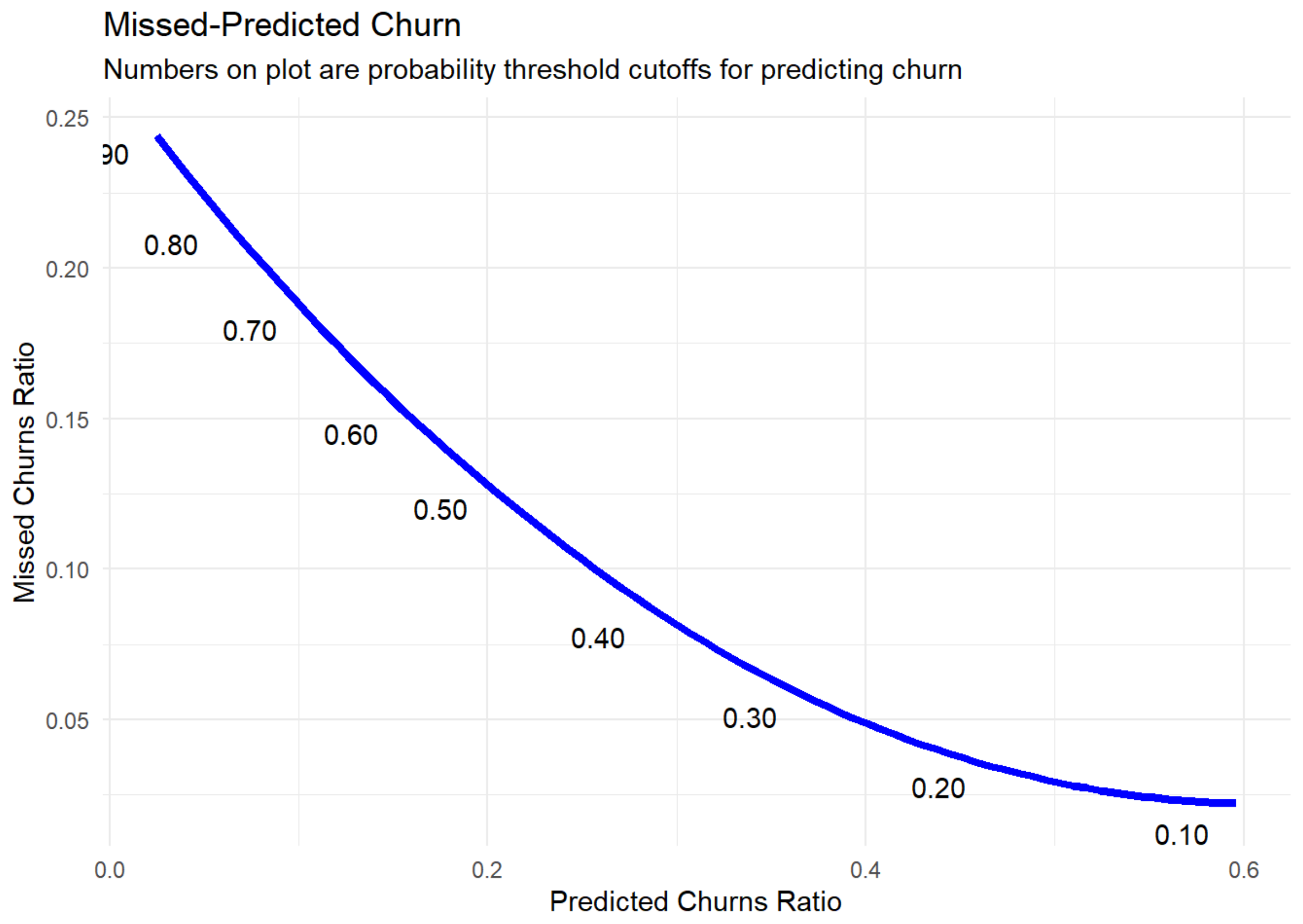
(#tab:table 2)Performance Metrics Across Different Sampling Techniques

Metric	Original.Dataset	Random.Undersampling	Random.Oversampling
Accuracy Rate	0.80	0.75	0.75
F measure	0.80	0.75	0.76
Precision	0.85	0.91	0.90
Sensitivity	0.88	0.71	0.75

Our Findings

The Feature Importance plot lists the top ten variables for our churn prediction model, highlighting referrals, dependents, tenure, contracts, and internet service as the most important.





The plot above illustrates how adjusting the probability threshold for churn classification impacts the F1-score, missed churn ratio, and predicted churn ratio, aiding in the selection of an optimal threshold for balanced model performance.

$$\text{Cost} = \text{RetentionCost}(\text{PredictedChurn}) + \text{ChurnCost}(\text{MissedChurn})$$

Table 1: Model Evaluation Metrics

Model	Precision	Recall	F1 Score	Accuracy
Decision Tree	0.70	0.36	0.47	0.79
RF 50% threshold	0.70	0.36	0.47	0.79
RF 33% threshold	0.59	0.75	0.66	0.80

Geographical Churn Reasons

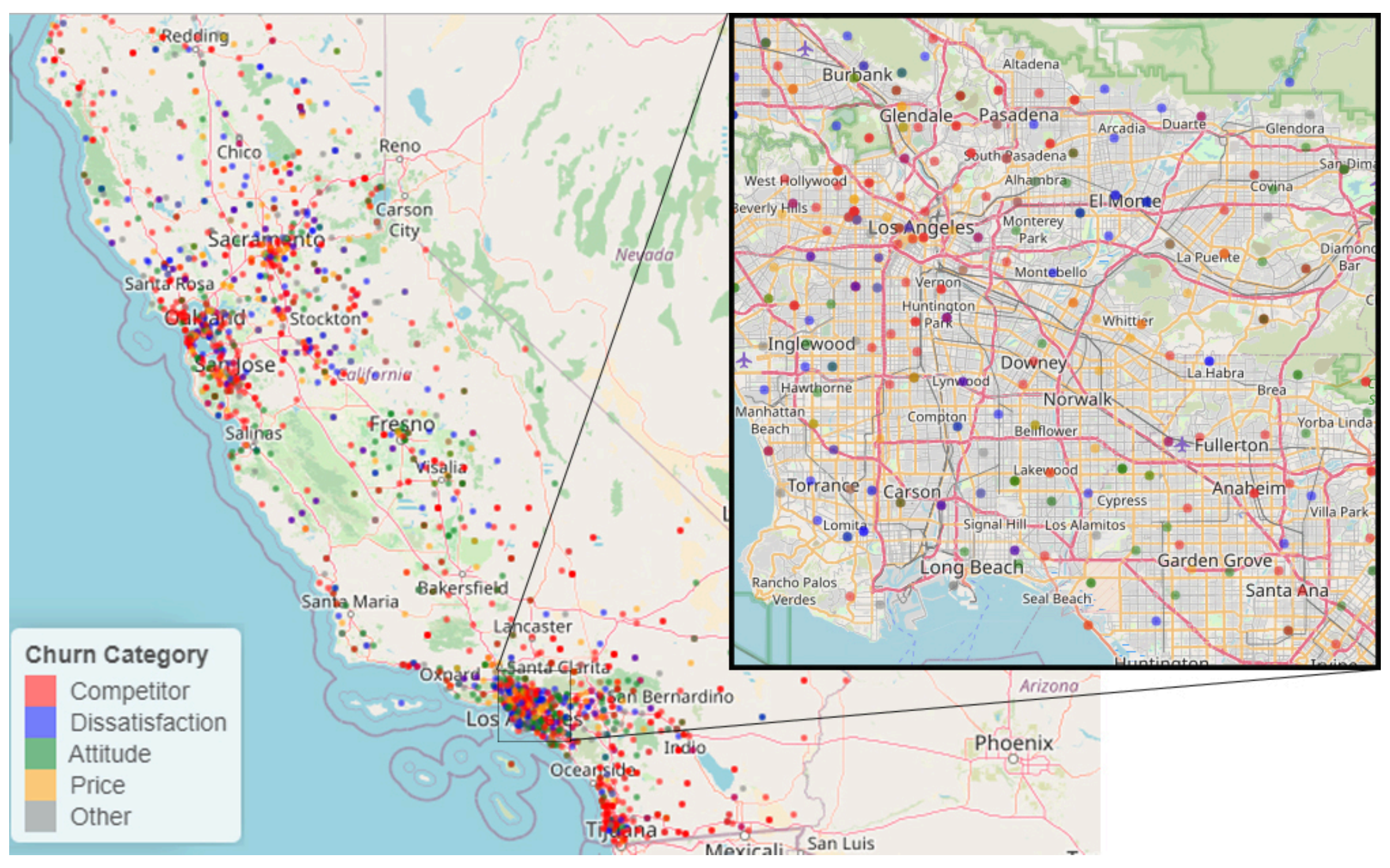


Figure 5: Map

The plot above shows that in Los Angeles, most customer churn is due to competitors and dissatisfaction.

Conclusion & Recommmendations

Our study emphasizes the crucial role of machine learning, specifically decision tree and random forest models, in accurately predicting customer churn and optimizing customer retention. Key recommendations include:

- **Prioritizing relevant variables** in the construction of churn prediction models to enhance effectiveness.
- **Tuning probability thresholds** in models to tailor predictions to specific business needs, thereby improving their utility in reducing churn.

These strategies are vital for addressing churn prediction challenges in the telecommunications industry.

References

Ahmad, A. K., A. Jafar, and K. Aljoumaa. 2019. "Customer Churn Prediction in Telecom Using Machine Learning in Big Data Platform." *J Big Data* 6. https://doi.org/10.1186/s40537-019-0191-6.
Bansal, A. 2023. "Telecommunications Industry Customer Churn Dataset." https://www.kaggle.com/datasets/aadityabansalcodes/telecommunications-industry-customer-churn-dataset.
Beeharry, Y., and R. T. Fokone. 2021. "Hybrid Approach Using Machine Learning Algorithms for Customers' Churn Prediction in the Telecommunications Industry." *Concurrency and Computation* 34 (4). https://onlinelibrary-wiley-com.plymouth.idm.oclc.org/doi/full/10.1002/cpe.6627.
Bilisik, O. N., and D. T. Sarp. 2023. "Analysis of Customer Churn in Telecommunication Industry with Machine Learning Methods." *Journal of Science & Technology* 11 (4): 2185–2208. https://dergipark.org.tr/en/download/article-file/2205764.
Investopedia. 2023. "Churn Rate: What It Means, Examples, and Calculations." https://www.investopedia.com/terms/c/churnrate.asp.
Luck, I. 2023. "Churn Rate: How to Calculate Customer Churn [with Formula]." https://customergauge.com/blog/churn-rate.
Ly, Y. N. N., and D. V. T. Son. 2022. "Churn Prediction in Telecommunication Industry Using Kernel Support Vector Machines." https://doi.org/10.1371/journal.pone.0267935.
Shabankareh, M. J., A. Nazarian, A. Ranjbaran, and N. Seyyedamiri. 2021. "A Stacking-Based Data Mining Solution to Customer Churn Prediction." *Journal of Relationship Marketing* 21 (2): 124–47. https://www-tandfonline-com.plymouth.idm.oclc.org/doi/full/10.1080/15332667.2021.1889743.
Tessitore, S. 2023. "What's the Average Churn Rate by Industry?" https://customergauge.com/blog/average-churn-rate-by-industry.

UNIVERSITY OF PLYMOUTH

posterdown