Data Challenge - Bank Marketing

Ameya Diwan **DSA Candidate - Ethical AI**

Table of Contents

Background

Data Cleaning

Exploratory Analysis and Insights

Data Preparation

Model Evaluation

Recommendations

Background

Objective

- Assist a banking institution to facilitate their marketing efforts, by predicting clients who will be more likely to subscribe to a term deposit product, based on historical campaign data
  - By uncovering valuable insights and delivering recommendations
  - By evaluating current model and potentially build a better predictive model for future campaigns

Increase revenue by boosting conversions, and decrease telemarketing costs by more efficient targeting

Dataset Overview

41188 records

Each record corresponds to every call made to the bank's clients

20 inputs

Client's socio-demographical info,
Data from previous contact with client and macro
Social and economic factors
(10 numeric, 10 categorical)

Target Variable

Indicates whether the client subscribed to the term deposit

HIGHLY IMBALANCED

(11.27% subscribed, 88.73% did not subscribe)

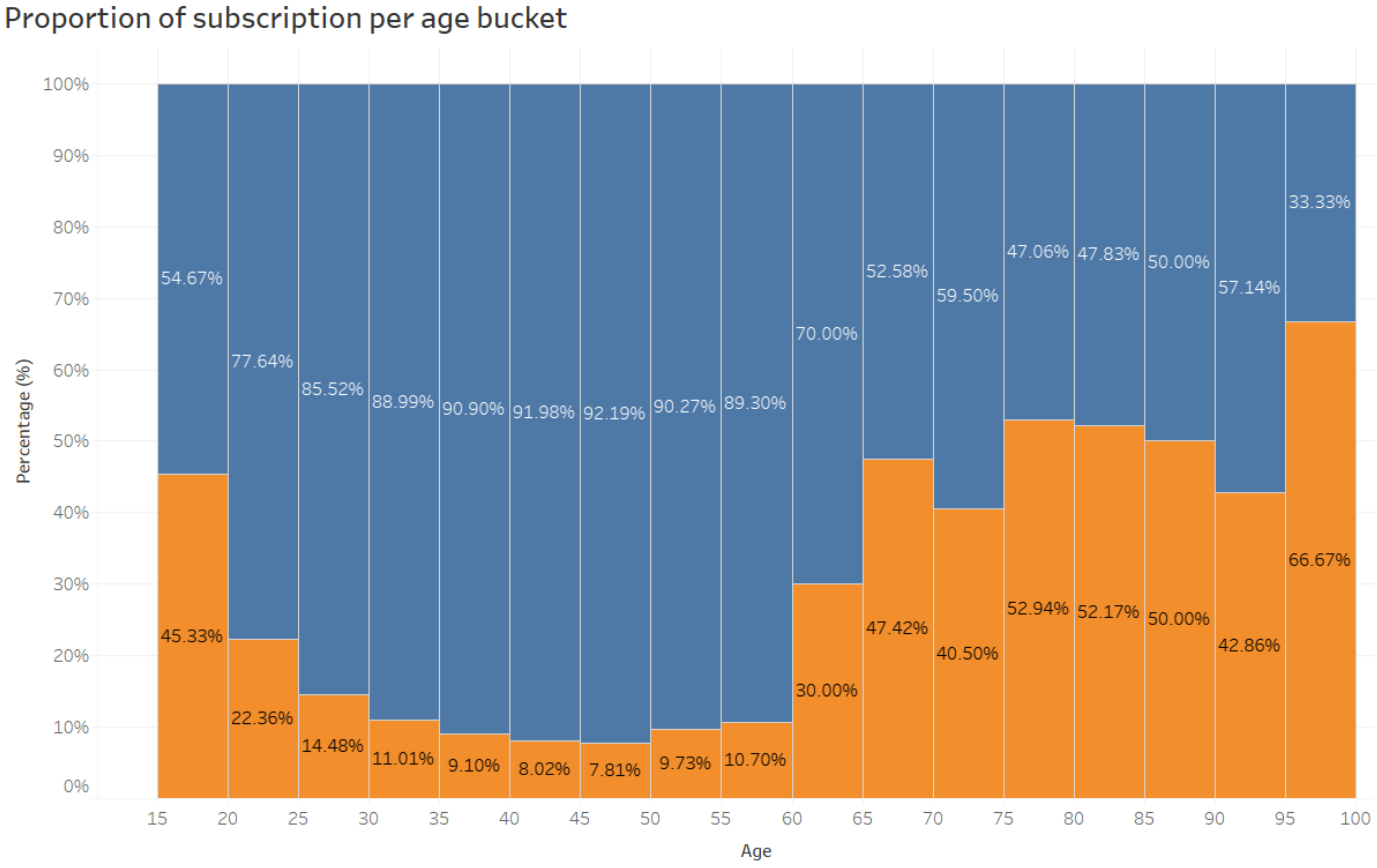
Data Cleaning

Cleaning Steps

- Removed duplicates
- Analyzed variables with unknown values present
- Addressed discrepancies in data with respect to previous marketing contacts
- Separated the call duration variable from the dataset

Exploratory Analysis

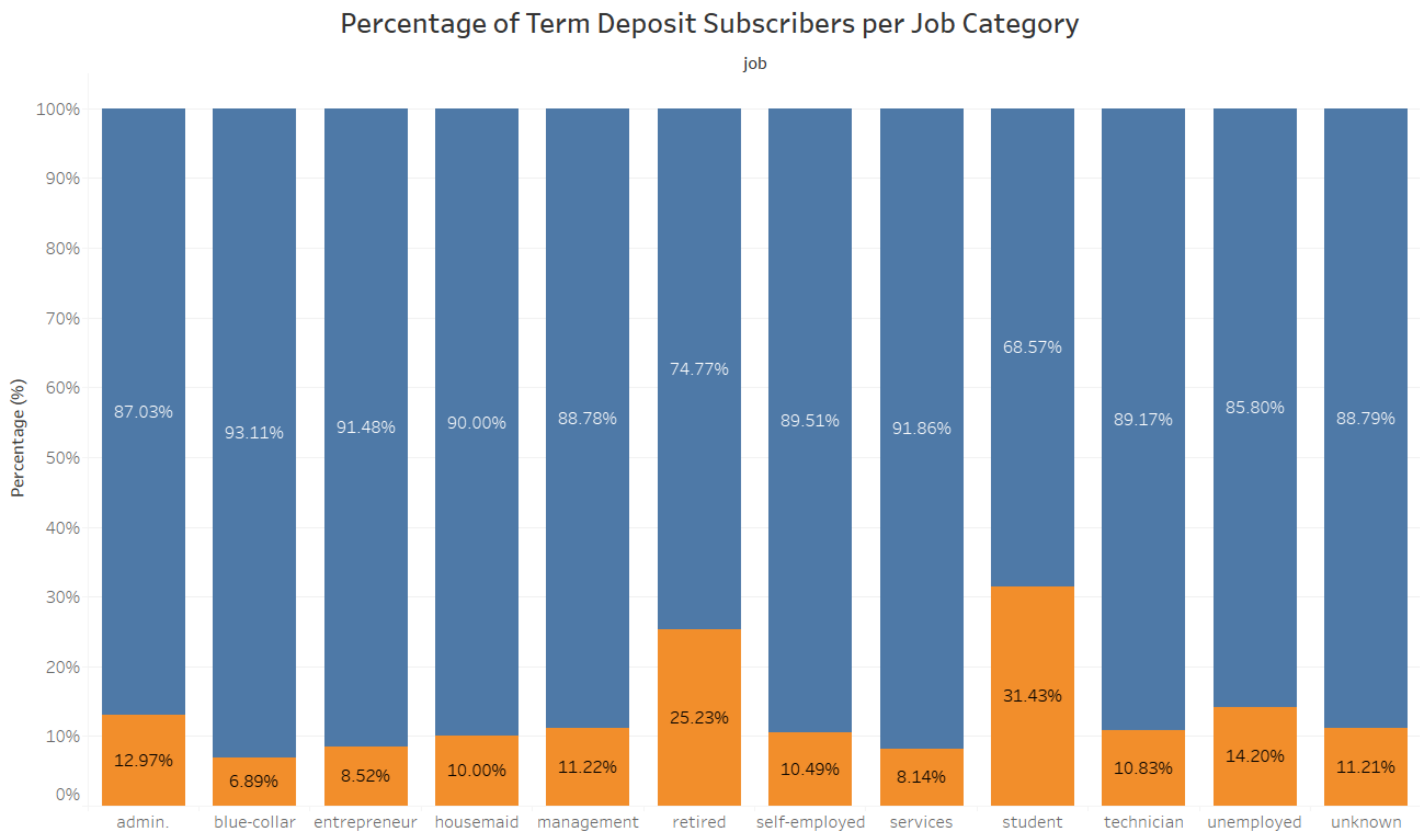
Insights in Focus



- **Majority of targeted clients are within ages 20 and 60, but interestingly, conversions are significantly higher in clients outside those ranges**

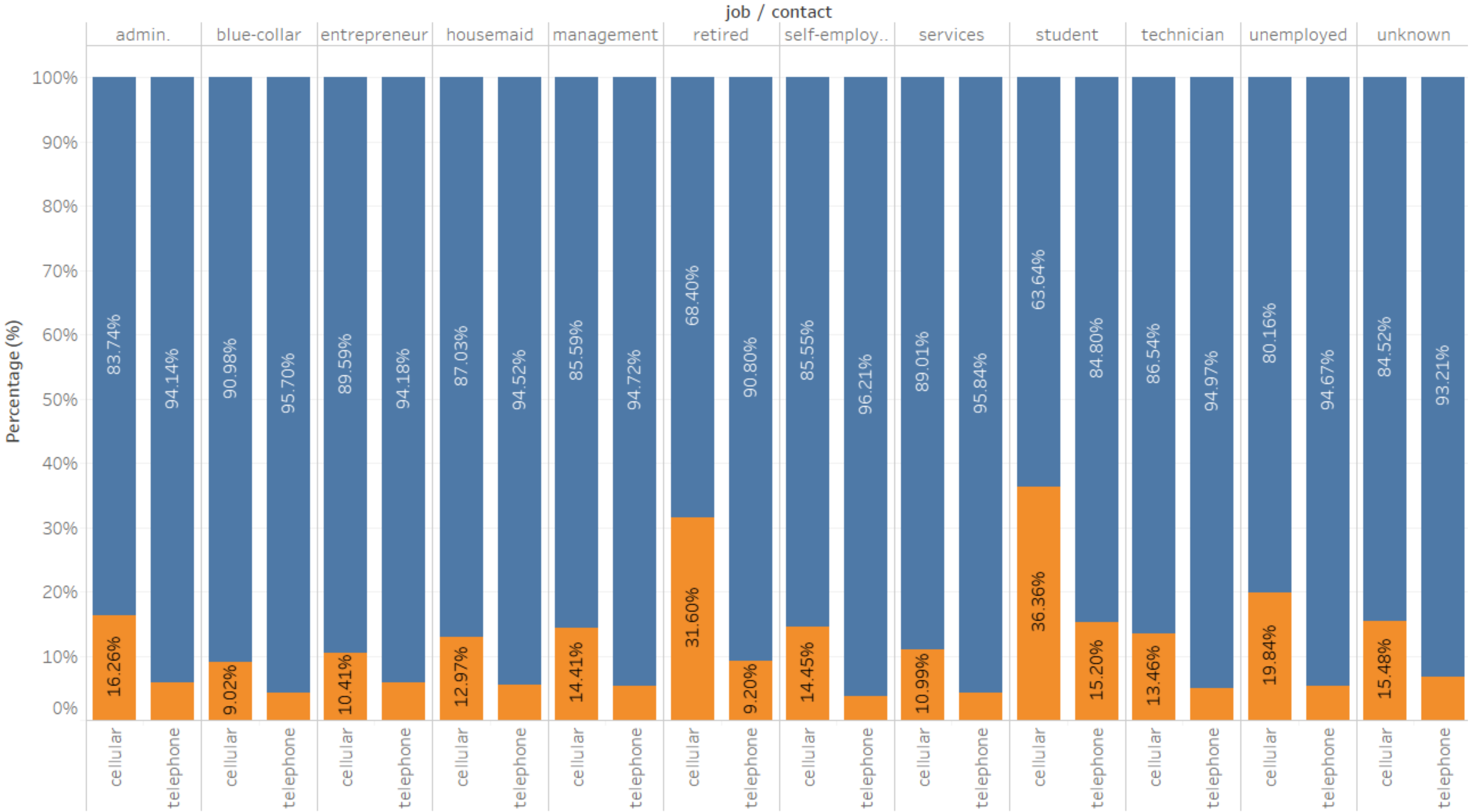
Insights in Focus

- Students and retired clients showed higher proportions subscribing compared to their working counterparts



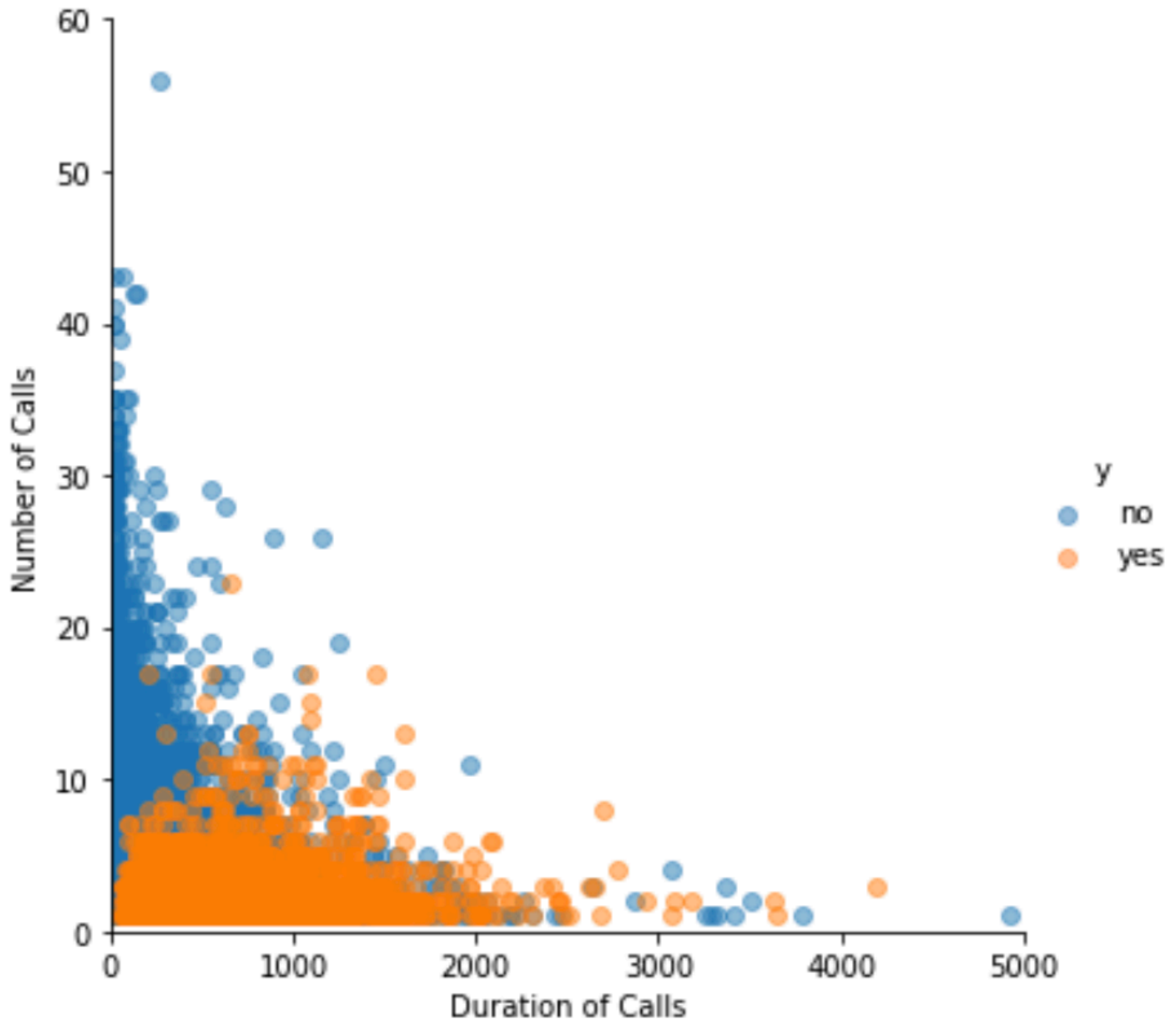
Insights in Focus

- Students and retired clients showed higher proportions subscribing compared to their working counterparts

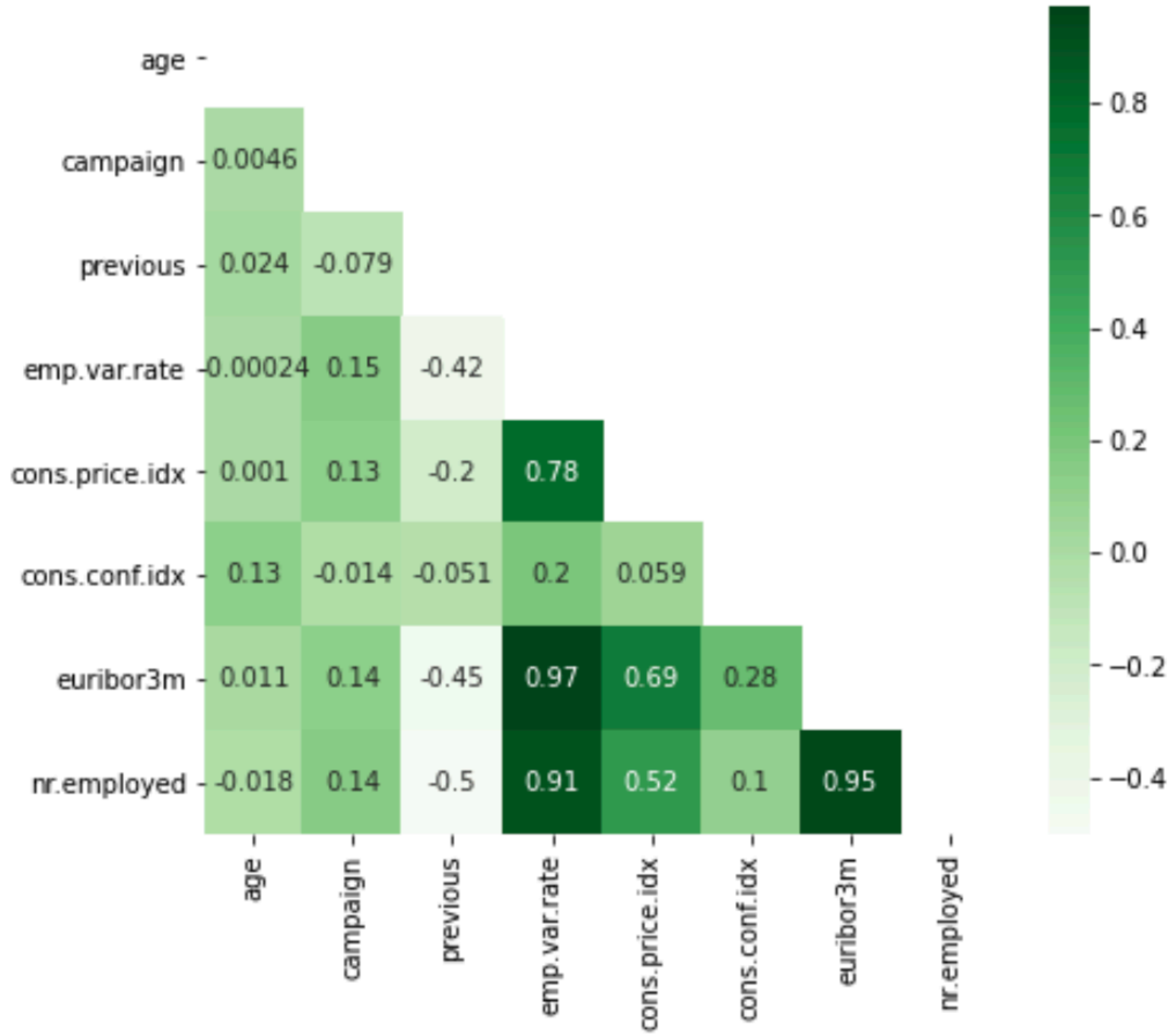


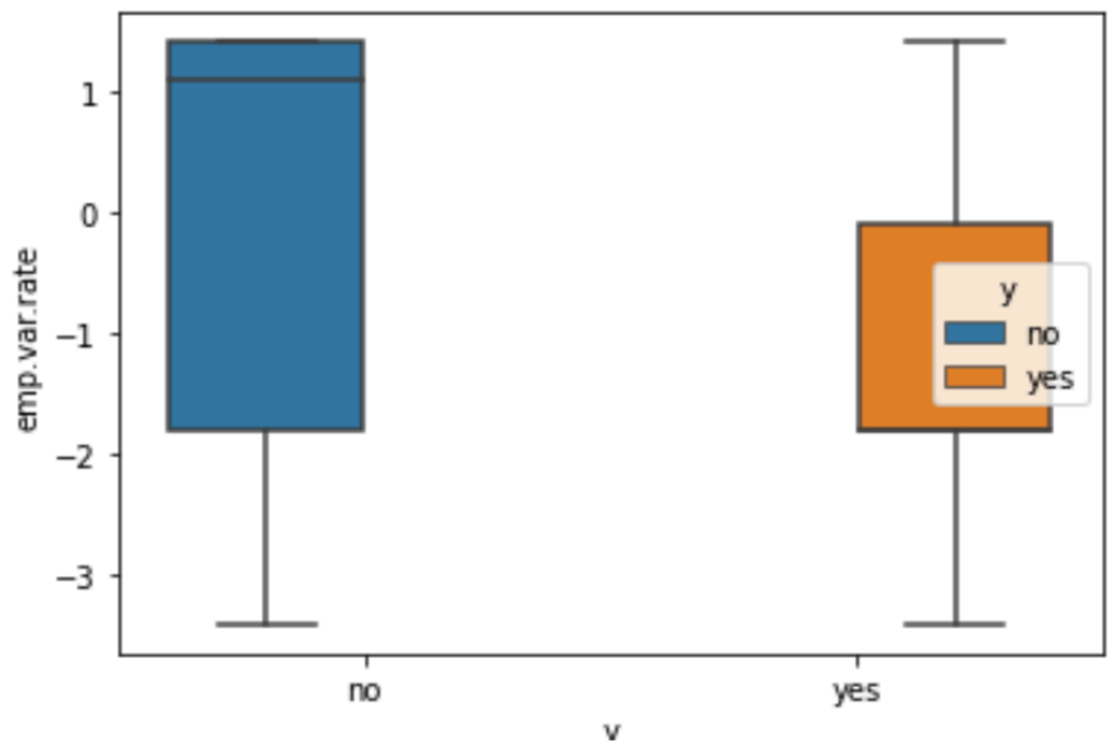
Insights in Focus

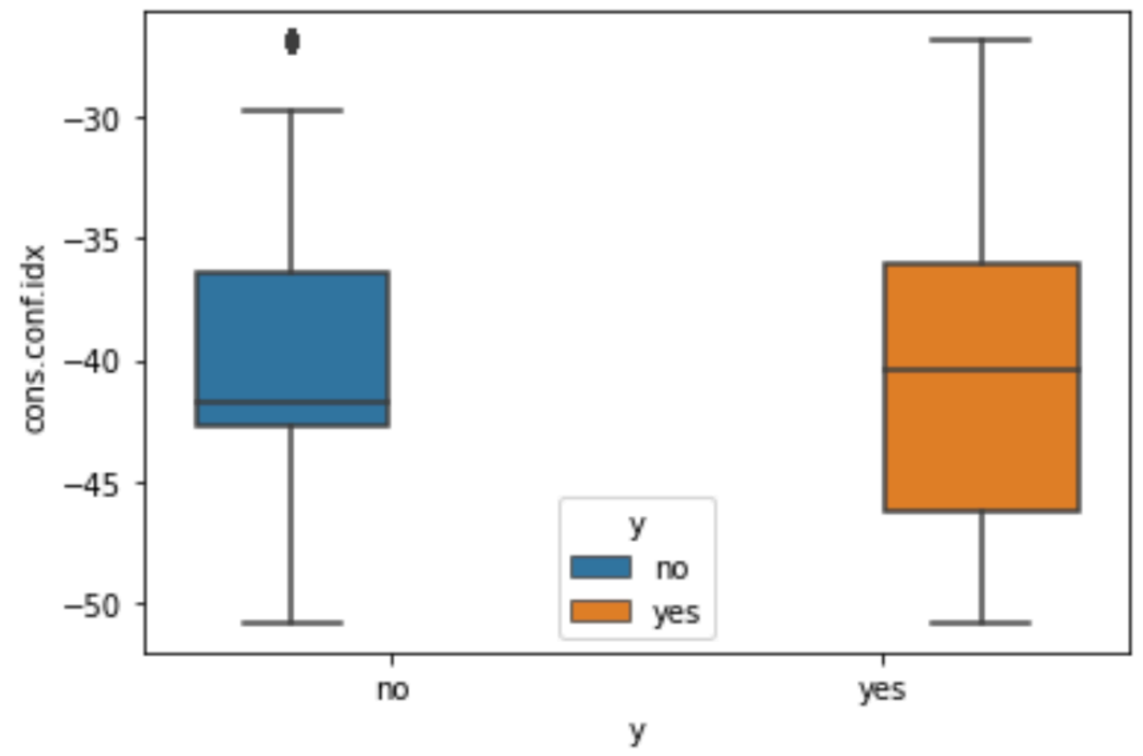
- **Conversions were optimized when clients were contacted multiple times but the total number of calls to each client was limited to not more than 10, and the total duration of calls was kept below 2500 minutes**

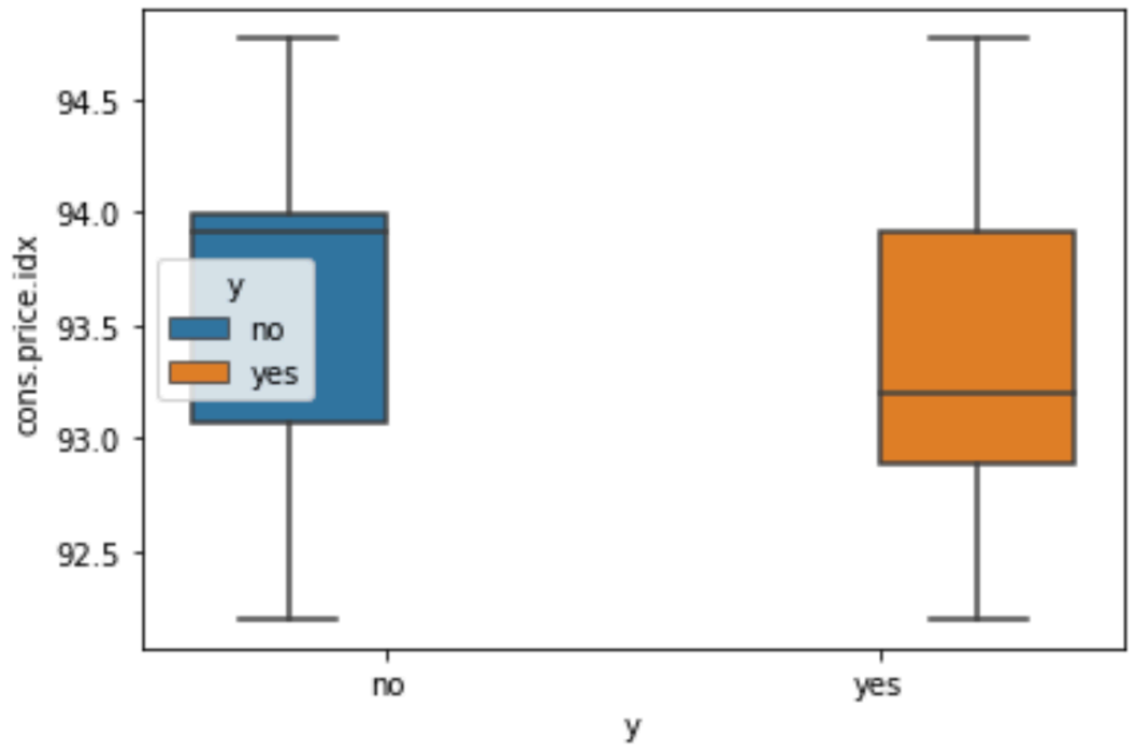


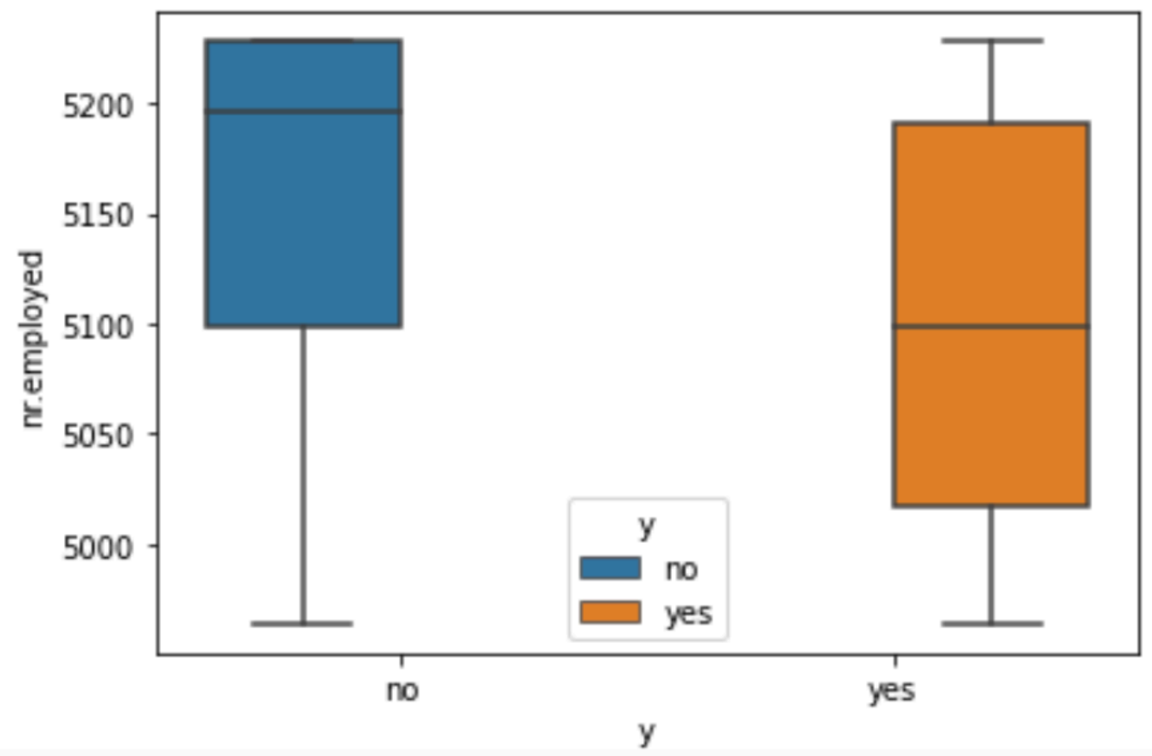
Insights in Focus



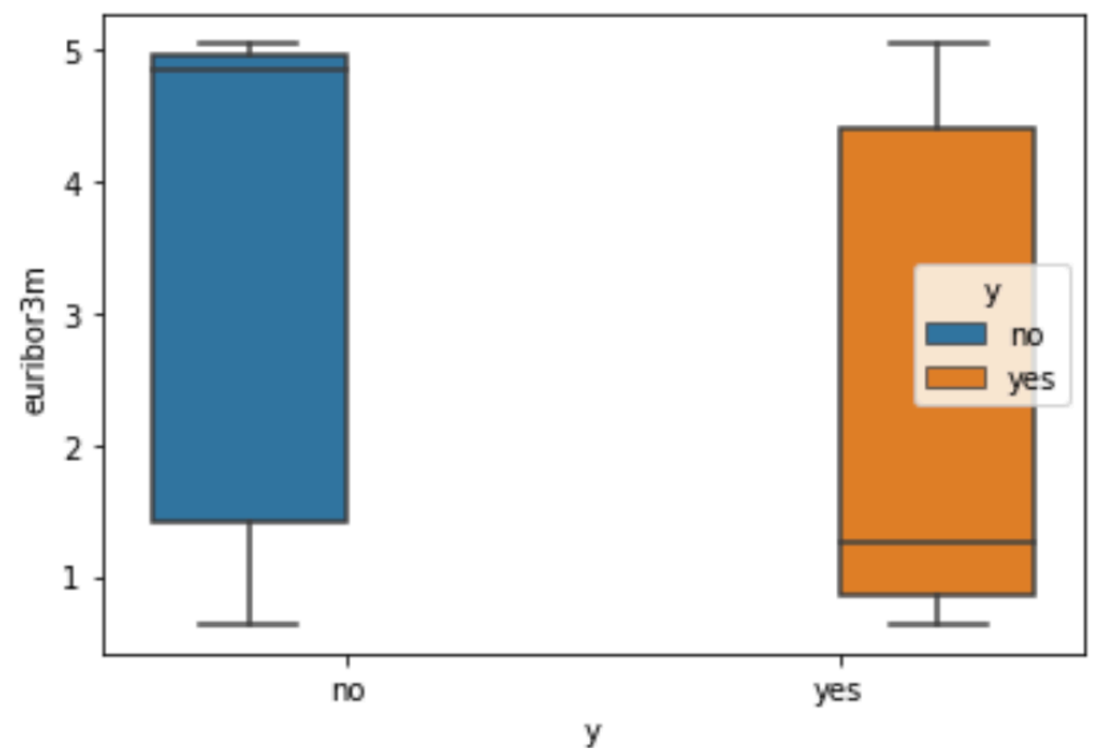








The socio-economic indicators are mostly correlated, and also point to patterns with term deposit subscriptions among contacted clients during that timeframe.



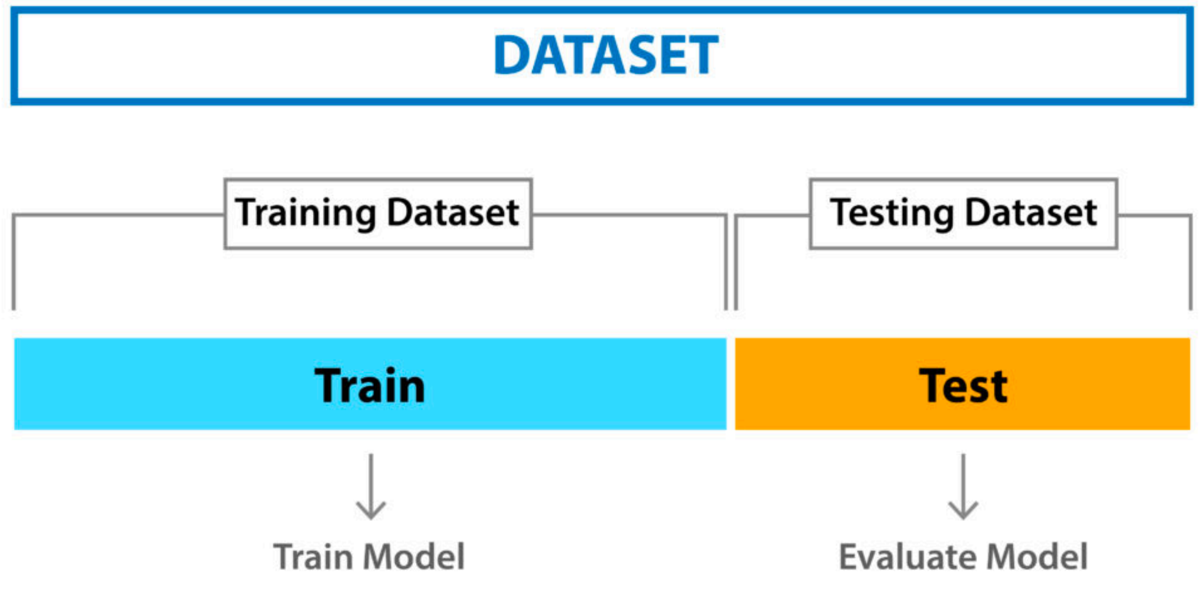
Data Preparation

Data Preparation

- Converted categorical columns to be represented numerically

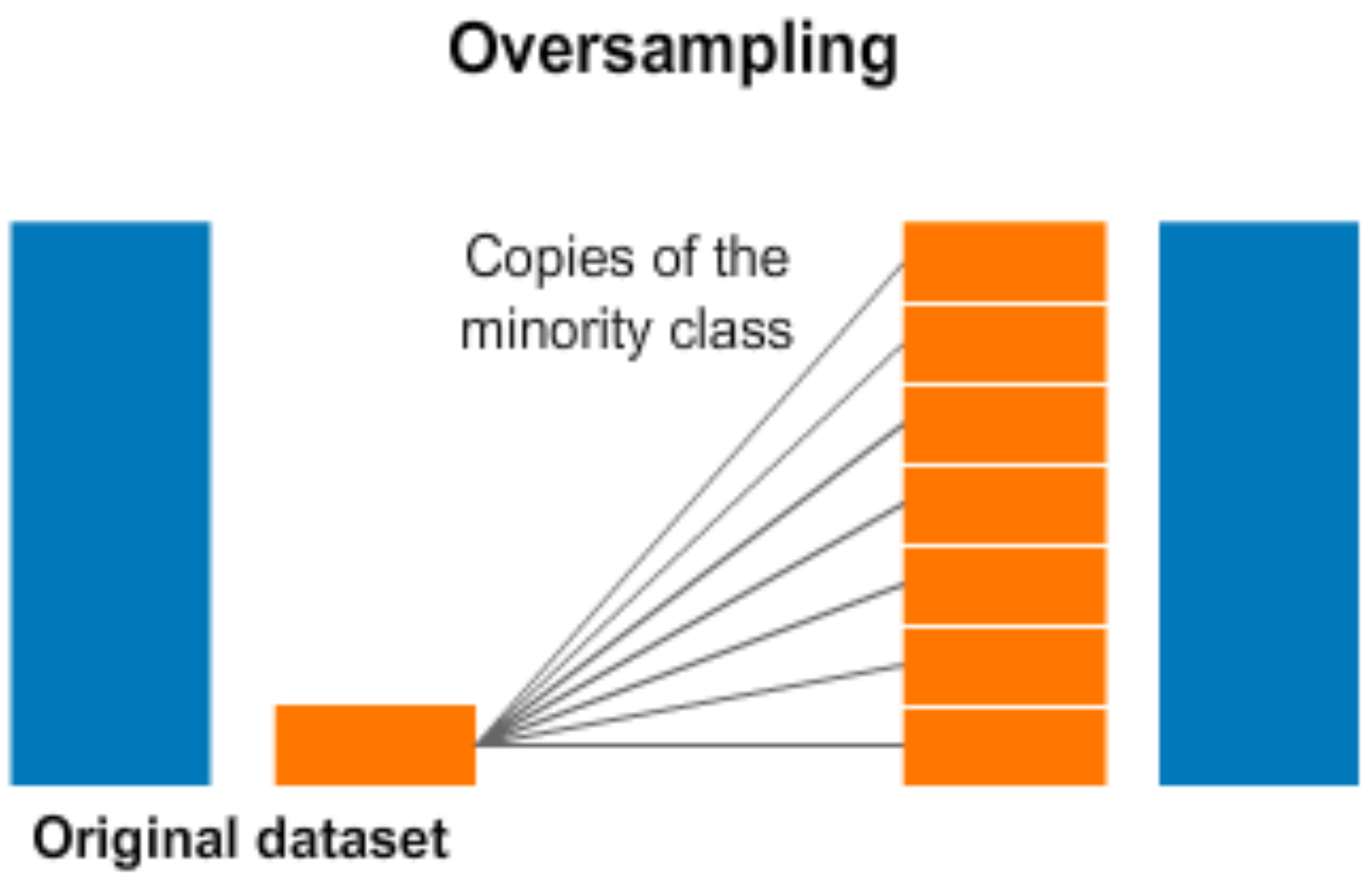
color	color_red	color_blue	color_green
red	1	0	0
green	0	0	1
blue	0	1	0
red	1	0	0

- Split data into sets for building and evaluating the model

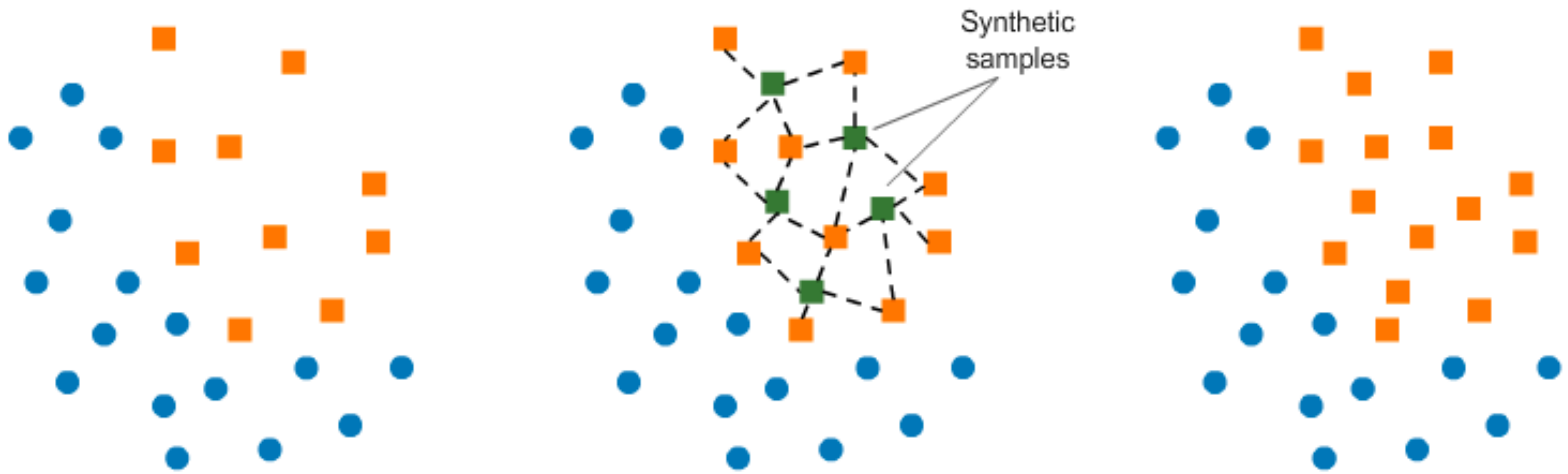


Data Preparation

- After initial model evaluation, we carry out synthetic multiplication of data-points in our training set based on data patterns in the training set itself, before training the model again



Source: kaggle.com



Source: kaggle.com

Model Setup and Evaluation

Model Setup Flow

- Define goal(s) and metrics of interest

- Decide candidate models

- Train model on training data segment

- Mix and match necessary parameters and tune model

- Evaluate performance on unseen test data and compare goal metrics

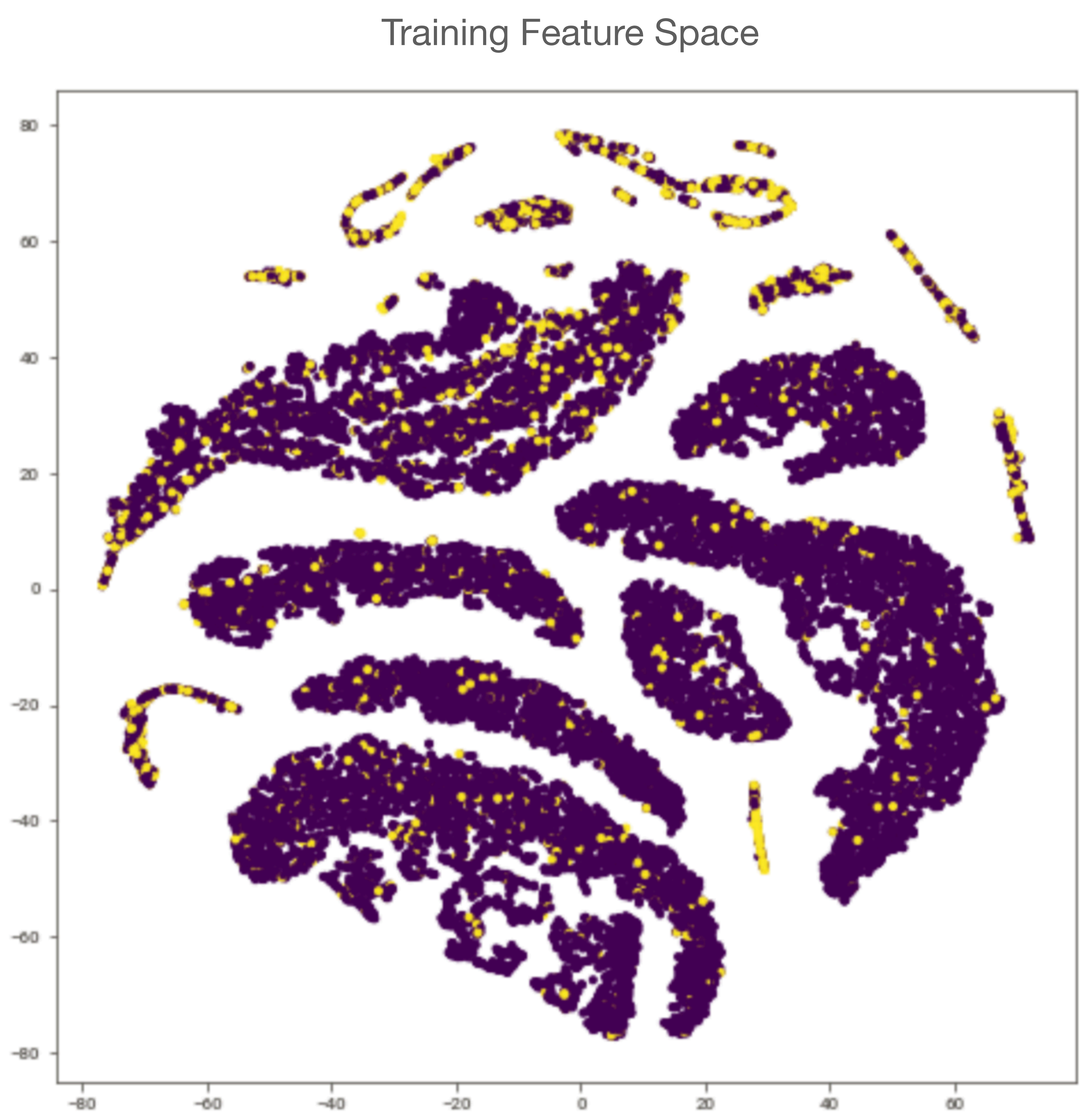
Evaluation Rubric

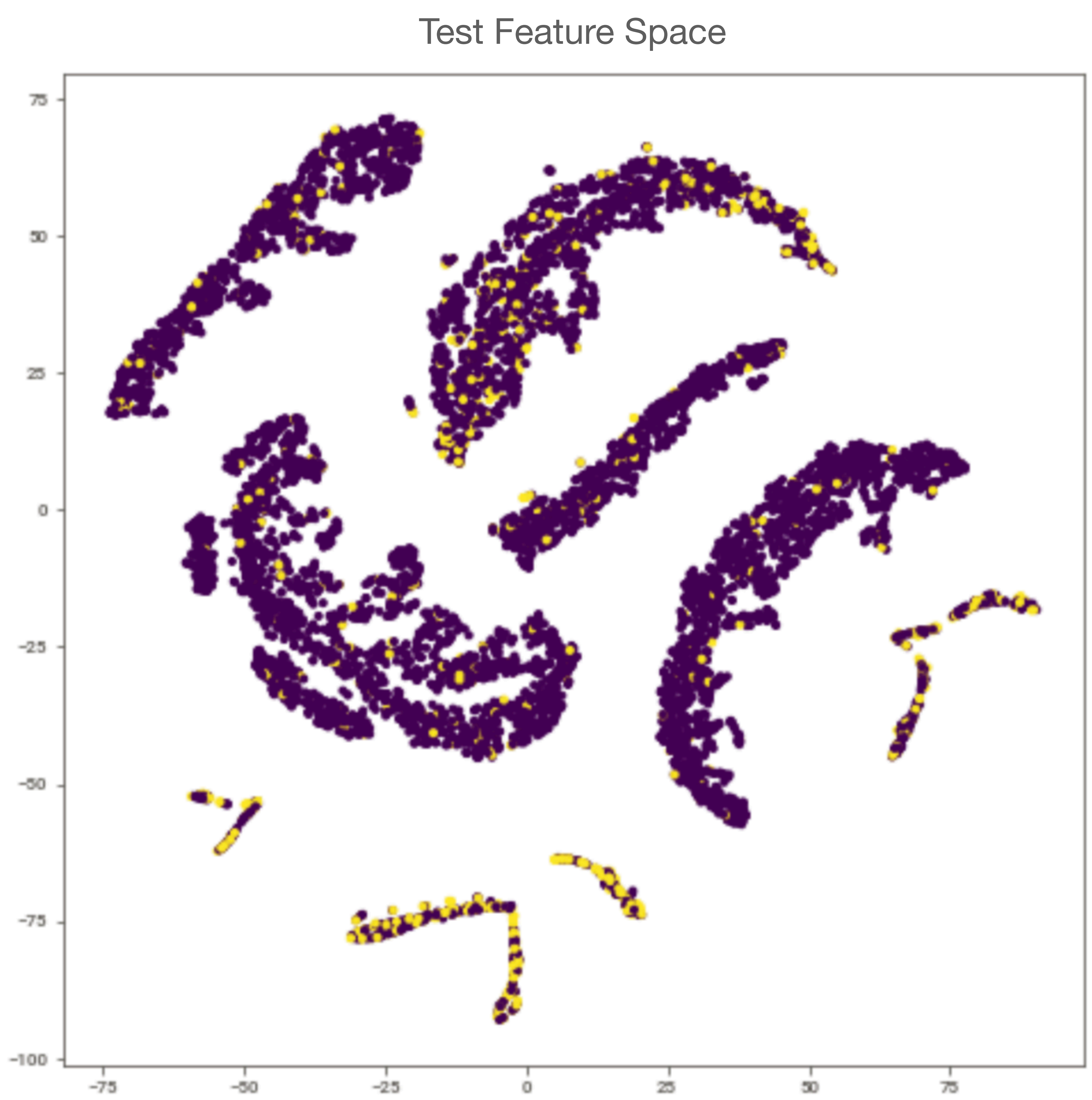
- Goal is to maximize the number of subscribers predicted correctly

- Correctly identifying unlikely subscribers is secondary

- Consider metrics to evaluate how well our model performs and identifies subscribers:
  - Recall : Total proportion of subscribers correctly identified
  - Precision : Total proportion of predictions that are subscribers
  - ROC Score : Measure of model performance by comparing True Positive Rate and False Positive Rate

Visualized Feature Space





Model Consideration

- We consider RandomForest and XGBoost models as primary options
  - High performance; ability to accommodate complex non-linear interactions within data
  - Highly tolerance to noise
  - Feature selection capabilities

Model Evaluation

Balanced data - Final Test data scores

- XGBoost model **successfully identifies 53% of the converting clients**
  vs 24% in the case of Random Forest

- Healthy balance between recall and precision for XGBoost
  vs imbalanced for Random Forest

- XGBoost has a better ROC of 0.792
  vs 0.783 in the case of Random Forest

Evaluation on Test data

	Random Forest	XGBoost
Precision	**0.58**	0.48
Recall	0.24	**0.53**
ROC AUC	0.783	**0.792**

Base Model Performance (Current)

- We have the current model's likelihood probability to be a subscriber, for each record in the dataset

- Focus will be on model's Recall, Precision and ROC metrics

Evaluation on entire data set

	Base Model (Original)
Precision	0.098
Recall	0.854
ROC AUC	0.235

Performance Evaluation

Full Dataset

- Both newly developed models significantly outperform the base model predictions overall

- Random Forest has highest score on the overall dataset but underperforms XGBoost on unseen test data, as it “over-learns” on the training data

Evaluation on entire dataset

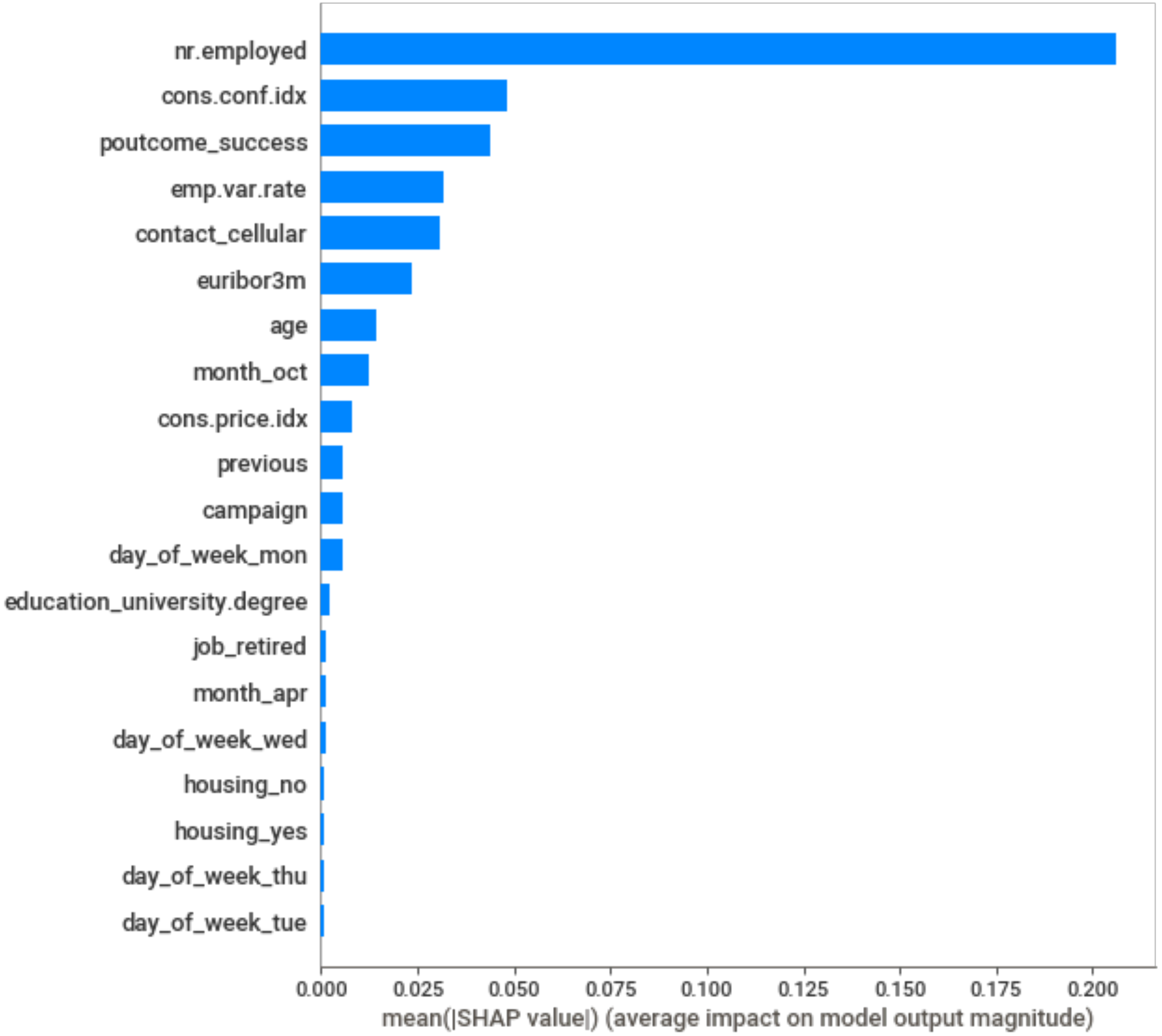
	Random Forest	XGBoost	Base Model (Original)
Precision	**0.58**	0.48	0.098
Recall	0.24	0.53	**0.854**
ROC AUC	0.783	**0.792**	0.235

Significant Features

XGBoost Model

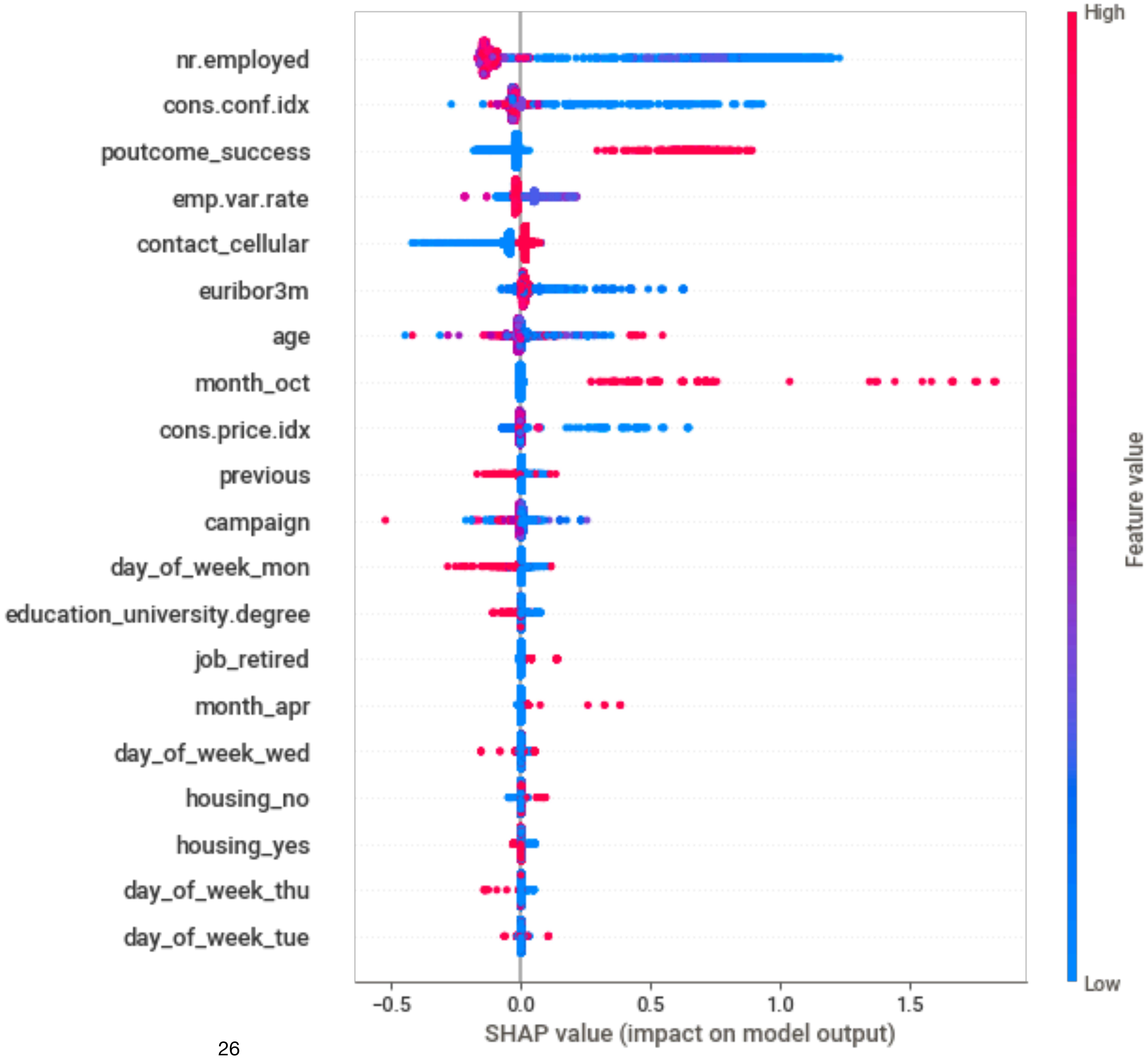
- All socio-economic variables play an important role in making predictions for the model

- Having a previous successful outcome in a marketing campaign, age and mode of the call are also all important factors



Relationship of Important Features with Target Variable

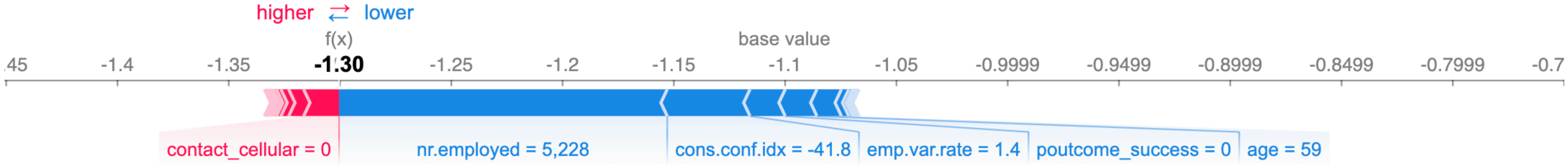
XGBoost Model



Explaining Predictions Locally

XGBoost Model

Prediction Explanation of the 42nd record in the training dataset



Recommendations and Next Steps

Recommendations

- Diversify targeting, especially in the case of ages and occupations; potentially offering term deposit services catered to the needs of younger and older clients, students and retired individuals

- Collaborate with economists and time marketing campaigns with respect to broader economic and employment conditions

- Prioritize customers who were a part of previous campaign efforts, and incorporate feedback scores from successful, as well as unsuccessful client conversions

- Optimize marketing efforts by limiting the number as well as duration of calls

Next Steps

- Continuously remodel and fine-tune the model to keep up with changing economic conditions and consumer behavior changes

- Evaluate models with dollar metrics involving true costs of similar campaigns

- Potentially involve more granular data-points bound to socio-economic nuances (e.g. BISG codes, micro-economic data by regions, etc.)

- Allocate resources for monitoring biases in the model predictions based on protected client information, to ensure ethical targeting of clients

Thank you!