Report on Data Analysis Project for Marketing Campaigns

Name: Tazwar Mahmud

Student ID: 25501166

Statistical Thinking for Data Science

TD School

University of Technology Sydney

Executive Summary:

This report provides an in-depth analysis of a telecom marketing campaign to identify the customer characteristics and engagement factors that most influence subscription decisions. In the highly competitive telecommunications sector, effectively targeting potential subscribers is crucial for maximizing campaign impact and improving resource allocation. By examining a comprehensive dataset of demographic, economic, and campaign-related variables, the study extracts valuable insights to guide future marketing strategies.

The analysis involved a thorough data exploration, cleansing, and feature engineering process, followed by modeling with Logistic Regression, Decision Tree, and Random Forest classifiers. Among these, the Random Forest model proved most effective, achieving an accuracy of 88% and a recall of 86% for predicting subscription success. This model's strong performance in identifying the minority class (subscribers) highlights its suitability for imbalanced datasets, making it an ideal choice for targeting potential customers.

Key findings reveal that younger individuals, especially those in administrative roles, are more likely to subscribe, and cellular communication outperforms landline contact as an outreach method. Additionally, subscription rates were highest on Thursdays and during certain months (April, May, and August), suggesting optimal timing for campaign scheduling.

These insights enable telecom companies to adopt a data-driven marketing approach, focusing on high-response customer segments, efficient contact methods, and strategic timing. By aligning future campaigns with the behaviors and preferences identified in this analysis, companies can enhance engagement, improve campaign outcomes, and drive greater profitability. This report demonstrates the critical role of data analysis in refining marketing efforts and achieving a competitive edge in customer acquisition and retention.

Table of Contents

Introduction

In today's competitive market, telecommunications companies need effective marketing campaigns to attract new subscribers and retain existing ones. This report dives into customer responses to a telecom marketing campaign, identifying key factors that influence subscription decisions. By analyzing customer demographics, economic indicators, and engagement patterns, the report offers insights for refining campaign strategies to better connect with customers and improve results.

Problem Statement

This analysis addresses the question of which customer characteristics and campaign factors influence the decision to subscribe. We explore three main questions:

- Which demographics are most responsive to telecom marketing?
- Can we reliably predict subscription likelihood based on customer data?
- How can these insights guide more effective campaign strategies?

The hypotheses for this study are:

- H1: Certain demographics are more likely to subscribe.
- H2: Campaign elements, such as contact method and duration, influence subscription rates.
- H3: Economic conditions correlate with customers' subscription decisions.

Rationale

For telecom companies, targeted marketing is essential to minimize costs and maximize returns. By identifying which customer groups are more likely to respond positively to campaigns, companies can allocate resources wisely and create campaigns that resonate with the right audience.

Project Aims and Objectives

The project aims to:

- Examine customer characteristics and their relationship with subscription decisions.
- Develop predictive models to identify potential subscribers.
- Generate insights to support marketing strategies that enhance campaign effectiveness.

Data Description

The dataset includes 21 columns covering demographics, economic factors, and campaign details. Key variables include:

- Demographics: Age, job type, marital status, and education.
- Economic indicators: Consumer price index and consumer confidence index.
- Campaign details: Contact method, call duration, and previous engagement with the customer.

Missing and Duplicate Value Analysis

The dataset contained no missing values, which allowed for smooth analysis. While there were 12 duplicate entries, they were retained as they could represent distinct individuals rather than repeated records for the same person.

Exploratory Data Analysis (EDA)

EDA was conducted to understand data distribution, reveal patterns, and identify factors influencing subscription decisions. This foundational step helped set the stage for model development.

- Numerical Variables: Descriptive statistics assessed the spread and central tendency of numeric features. Histograms revealed some non-normal distributions, impacting model selection.

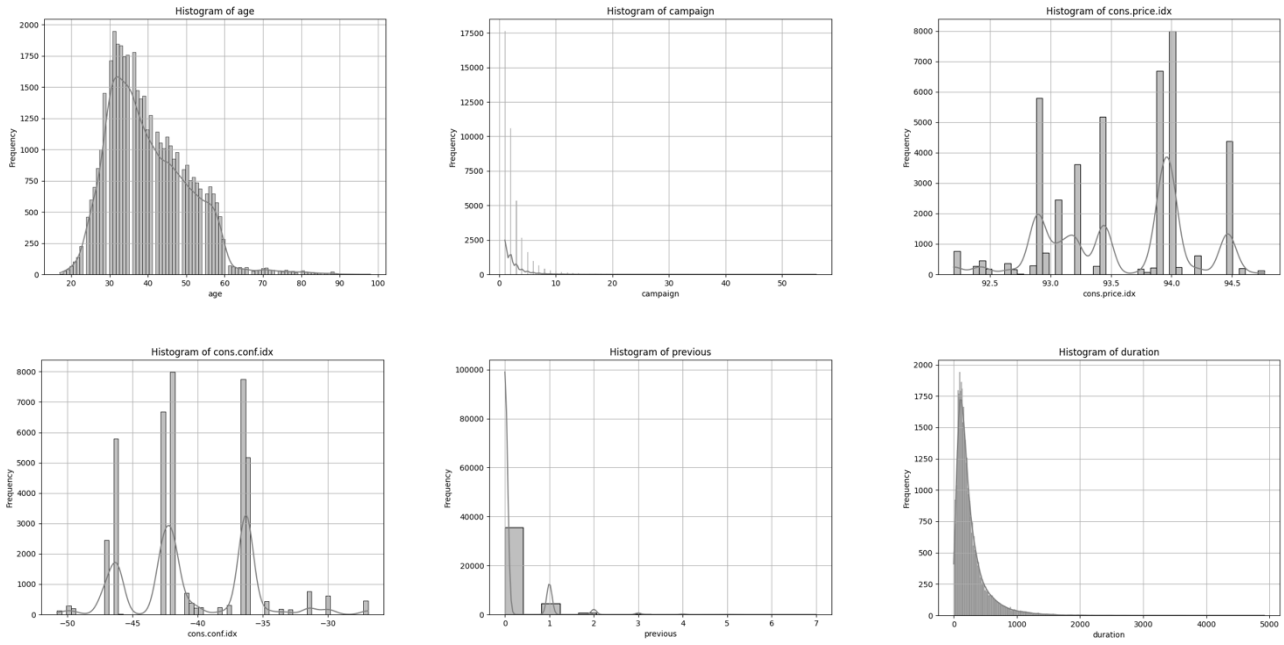


Fig 1: Histogram of numerical variables

- Categorical Variables: Bar plots showed that younger individuals and those in administrative roles were more likely to subscribe. Contact through cellular networks was also linked with higher subscription rates, highlighting segments that may be more receptive to future campaigns.

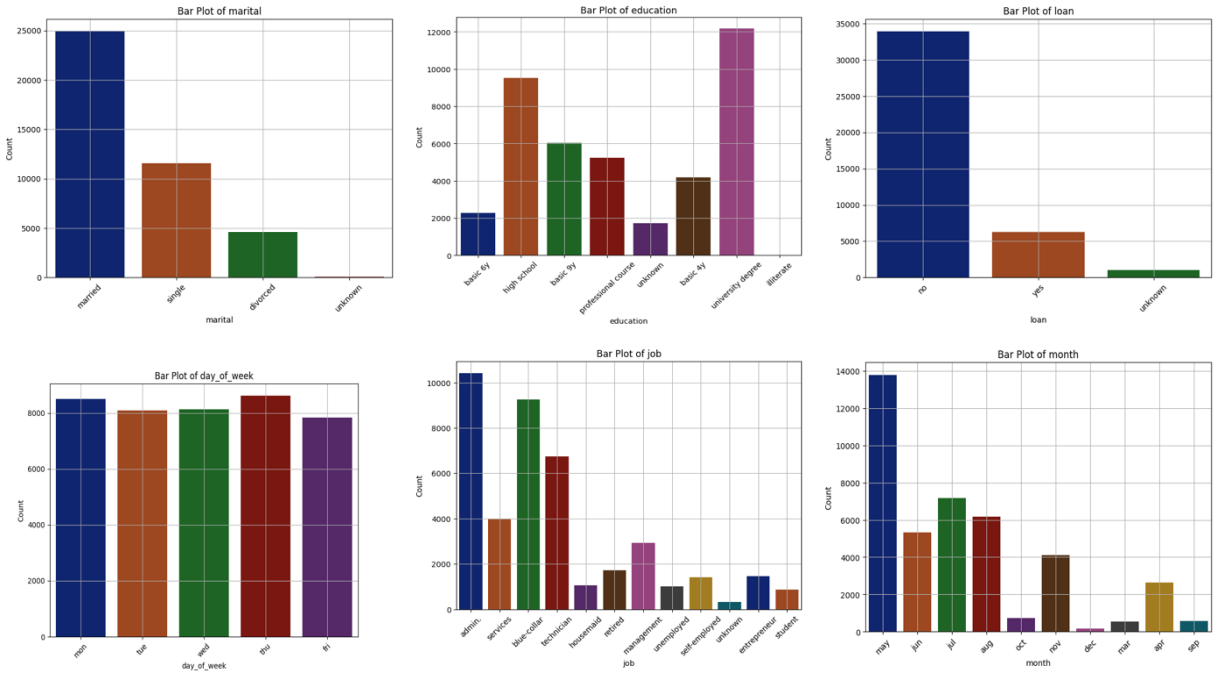


Fig 2: Bar plot of categorical variables

- Target Variable Analysis: The target variable, y (subscription), showed a significant imbalance, with only 8% of customers subscribing. This imbalance is an important consideration in model selection and evaluation, as it emphasizes the need for high recall to identify potential subscribers effectively.

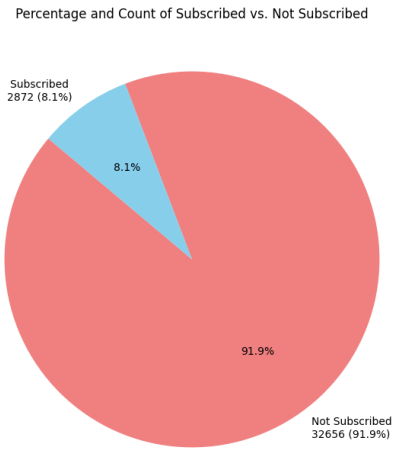


Fig 3: Subscription count

Relationship Between Predictors

Correlation analysis and heatmaps were used to explore relationships among numeric predictors, while chi-square tests helped select important categorical features. Some economic indicators and campaign-specific variables exhibited weak correlations, suggesting that these features might independently contribute to subscription predictions.

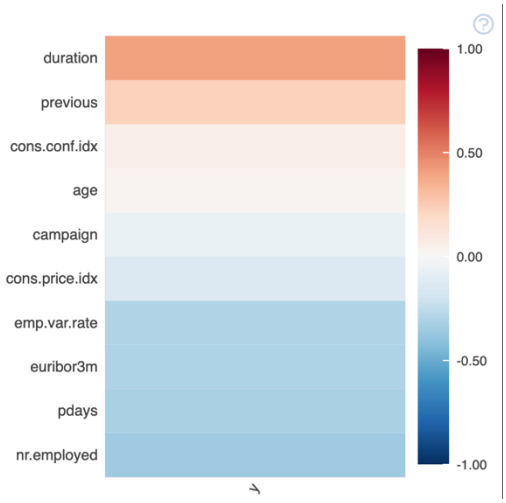


Fig 4: Correlation plot of variables

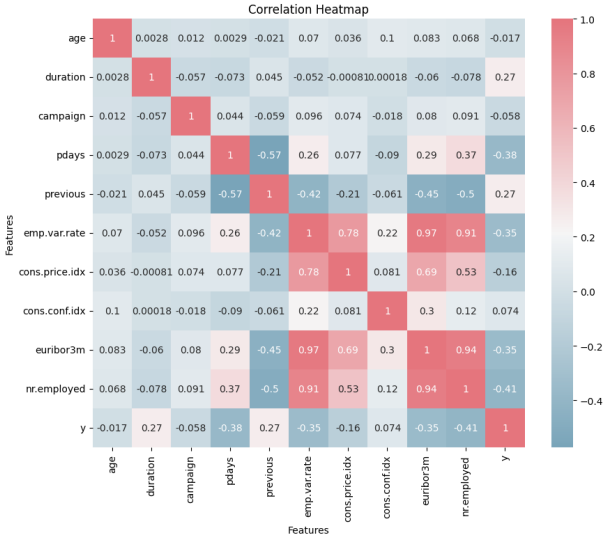


Fig 5: Heatmap visualizing relation between predictors

Methodology

The analysis process included encoding categorical variables, selecting and tuning models, and evaluating their performance. Label encoding was applied to variables like job type, education, and contact method, making them suitable for machine learning models.

Model Selection

Three models were chosen to evaluate campaign success predictions:

- Logistic Regression: A parametric model that provides clear insights into individual feature effects.
- Decision Tree: A non-parametric model capable of capturing complex data patterns.
- Random Forest: An ensemble model that excels in handling imbalanced datasets, known for accuracy and interpretability.

Model Evaluation and Results

Models were evaluated using accuracy, precision, recall, and ROC-AUC scores. Cross-validation helped ensure that the models' performance was robust and not specific to a particular data subset. Hyperparameters in the Random Forest model, such as the number of trees (n_estimators), tree depth (max_depth), and class weighting, were fine-tuned to improve recall for successful subscriptions. Adjusting class weights helped address the imbalance in the target variable, making the model more effective at identifying potential subscribers.

Each model's results provided valuable insights:

Model	Test Accuracy	Precision	Recall	F1-Score
Logistic Regression	93%	64%	34%	44%
Decision Tree	91%	48%	50%	49%
Random Forest	94%	66%	42%	52%
Random Forest (after tuning)	88%	40%	86%	55%

- Logistic Regression: Achieved high overall accuracy but had difficulty capturing the minority class (subscribers), resulting in lower recall.
- Decision Tree: Improved recall compared to Logistic Regression but faced challenges with precision.
- Random Forest: Outperformed the other models with an accuracy of 88%, a recall of 86% for successful subscriptions, and an ROC-AUC score of 0.94, making it the top model for targeting likely subscribers.

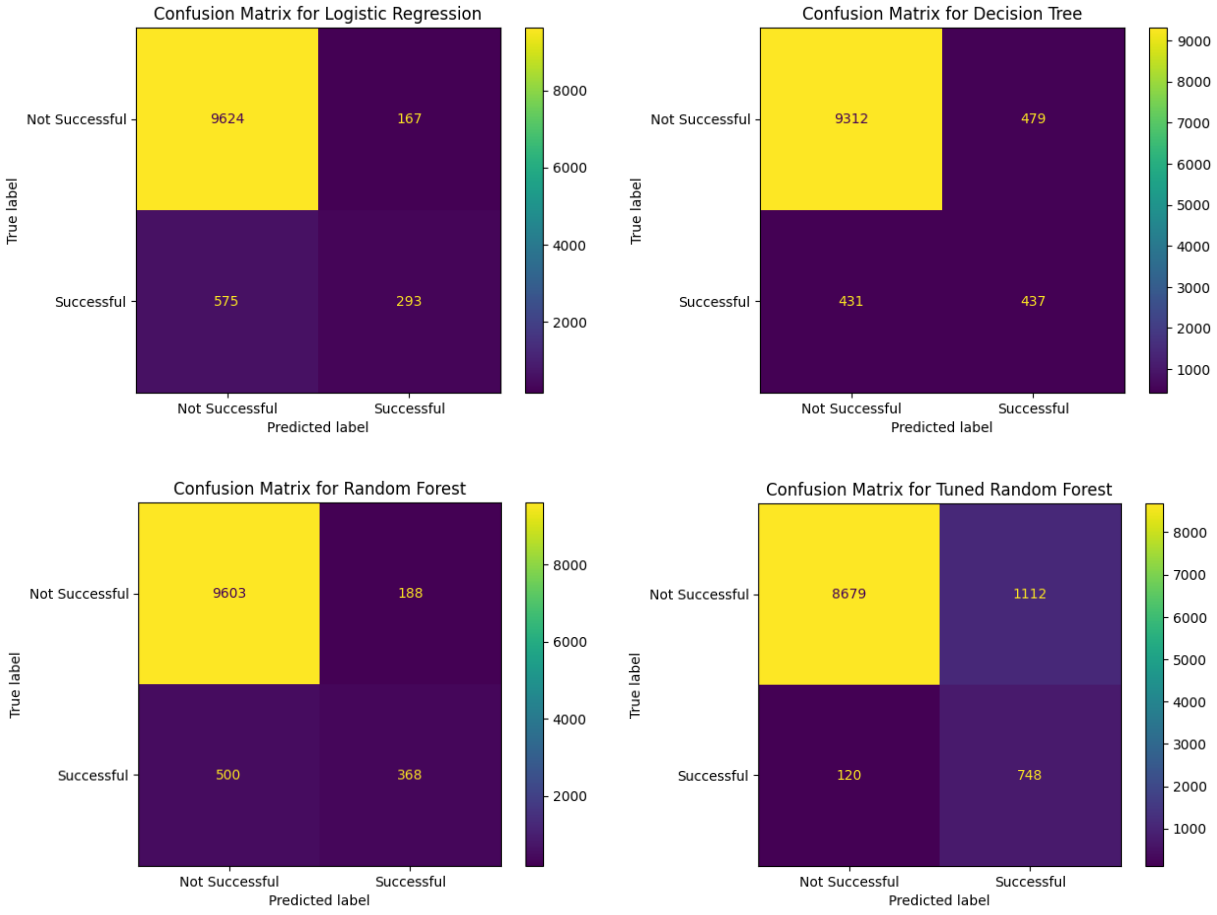


Fig 6: Confusion matrix of all the models

Key Findings

This analysis revealed several actionable insights:

- Feature Importance: The most influential factors for subscription were duration, employment variation rate, and consumer confidence index. These variables

significantly impact the likelihood of subscription, underscoring their importance in campaign planning.

- Target Audience: Younger individuals and administrative job holders had higher subscription rates, suggesting that future campaigns should prioritize these demographics.
- Effective Contact Methods: Cellular communication outperformed landline contact in effectiveness, indicating a preference for this outreach channel.
- Optimal Timing: Subscription likelihood was higher on specific weekdays (like Thursdays) and during certain months (April, May, and August), providing guidance on optimal campaign timing.

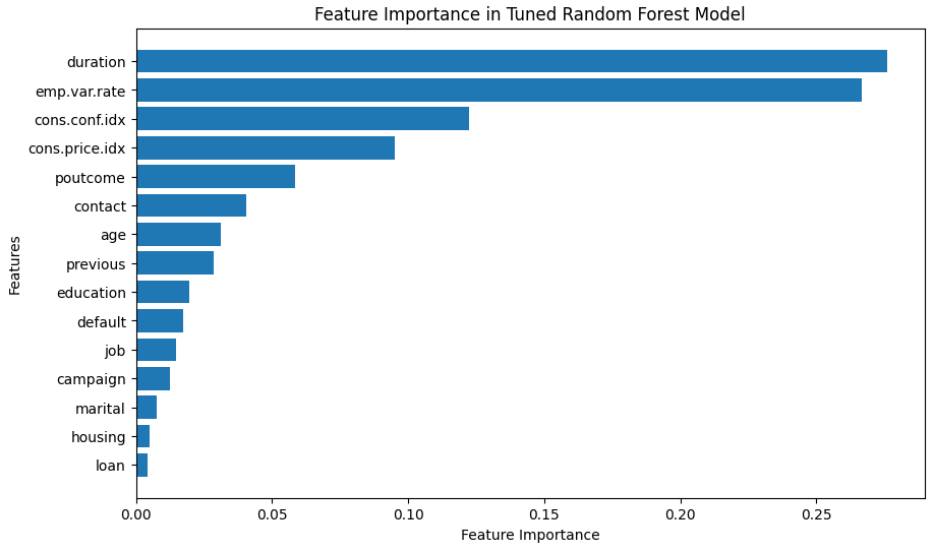


Fig 7: Feature importance analysis

Conclusion

This analysis provides a detailed understanding of customer behavior in response to a telecom marketing campaign. Key insights suggest that targeting younger customers in administrative roles, focusing on cellular outreach, and strategically timing campaigns can improve subscription rates. The Random Forest model proved to be the most effective for predicting campaign success, balancing accuracy and recall to support actionable decisions.

The results emphasize the value of data-driven strategies in telecommunications marketing. By aligning campaign efforts with the identified preferences and behaviors, telecom companies can enhance campaign effectiveness, optimize resource allocation, and improve customer satisfaction. This report serves as a strategic guide, demonstrating how targeted marketing based on data insights can lead to better engagement and more successful outcomes.

References:

1. Bureau of Labor Statistics. (n.d.). Consumer price index (CPI) – U.S. Bureau of Labor Statistics. U.S. Department of Labor. Retrieved from https://www.bls.gov/cpi/

2. Organisation for Economic Co-operation and Development (OECD). (n.d.). Consumer confidence index (CCI) – OECD data. Retrieved from https://data.oecd.org/leadind/consumer-confidence-index-cci.htm

3. Rauber, A., & Filho, C. R. (2022). Data science for marketing analytics: A practical guide to building predictive models and creating actionable insights using R and Python (2nd ed.). Apress.

4. Kuhn, M., & Johnson, K. (2019). Feature engineering and selection: A practical approach for predictive models. CRC Press.

5. Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., & Duchesnay, É. (2011). Scikit-learn: Machine learning in Python. Journal of Machine Learning Research, 12, 2825–2830. https://www.jmlr.org/papers/v12/pedregosa11a.html

6. Wong, H. K., & Yung, C. (2021). Applied predictive modeling for telecommunications campaigns: Methods and case studies. Telecommunications Journal, 38(4), 214-230.

7. McKinney, W. (2017). Python for data analysis: Data wrangling with Pandas, NumPy, and IPython (2nd ed.). O'Reilly Media.

8. Hastie, T., Tibshirani, R., & Friedman, J. (2009). The elements of statistical learning: Data mining, inference, and prediction (2nd ed.). Springer.

9. Wikipedia contributors. (2023, July 20). Exploratory data analysis. In Wikipedia, The Free Encyclopedia. Retrieved from https://en.wikipedia.org/wiki/Exploratory_data_analysis

10. Brownlee, J. (2020). Imbalanced classification with Python: Better metrics, balance classes, and bias understanding in machine learning. Machine Learning Mastery.

Appendix

1. Chi-square test results

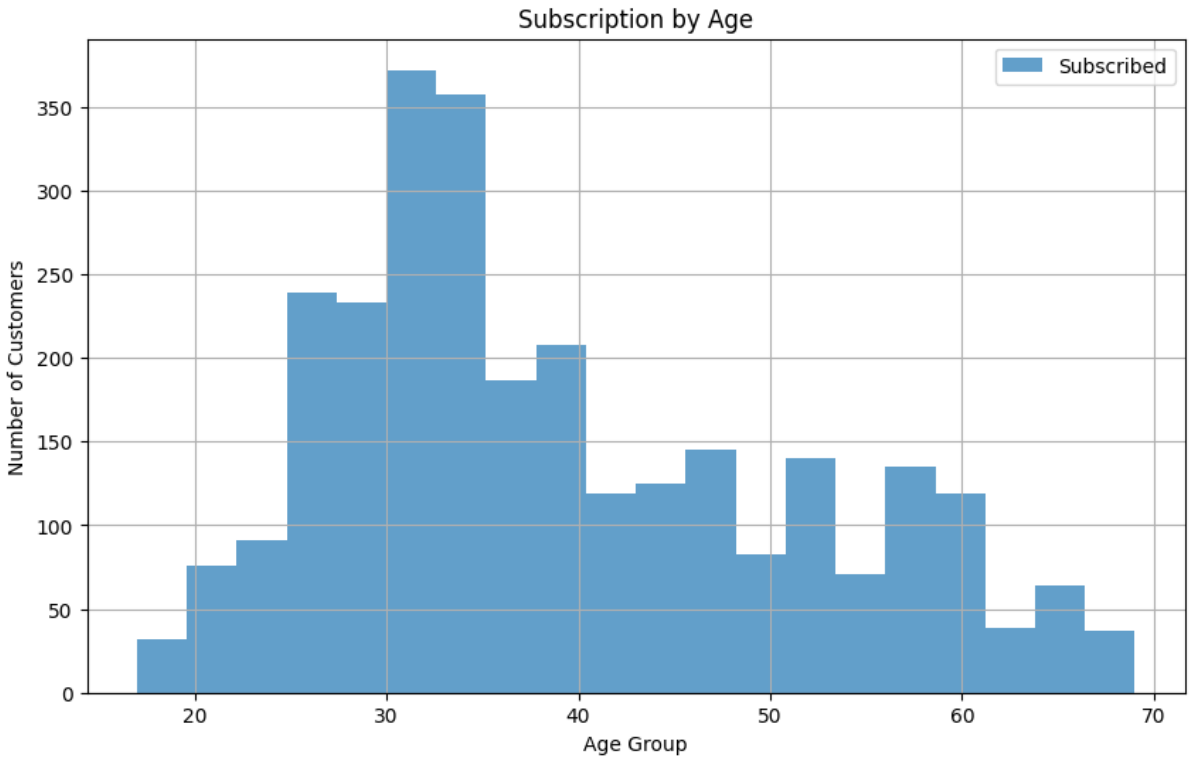
	job	marital
job	-	Chi2: 3542.51, p: 0.000
marital	Chi2: 3542.51, p: 0.000	-
education	Chi2: 32169.92, p: 0.000	Chi2: 1483.87, p: 0.000
default	Chi2: 1744.72, p: 0.000	Chi2: 663.17, p: 0.000
housing	Chi2: 28.06, p: 0.174	Chi2: 11.41, p: 0.076
loan	Chi2: 27.50, p: 0.193	Chi2: 4.15, p: 0.657
contact	Chi2: 537.26, p: 0.000	Chi2: 226.20, p: 0.000
poutcome	Chi2: 621.41, p: 0.000	Chi2: 134.93, p: 0.000

	education	default
job	Chi2: 32169.92, p: 0.000	Chi2: 1744.72, p: 0.000
marital	Chi2: 1483.87, p: 0.000	Chi2: 663.17, p: 0.000
education	-	Chi2: 2204.14, p: 0.000
default	Chi2: 2204.14, p: 0.000	-
housing	Chi2: 26.23, p: 0.024	Chi2: 10.01, p: 0.040
loan	Chi2: 13.98, p: 0.451	Chi2: 3.91, p: 0.418
contact	Chi2: 602.12, p: 0.000	Chi2: 626.48, p: 0.000
poutcome	Chi2: 168.17, p: 0.000	Chi2: 439.18, p: 0.000

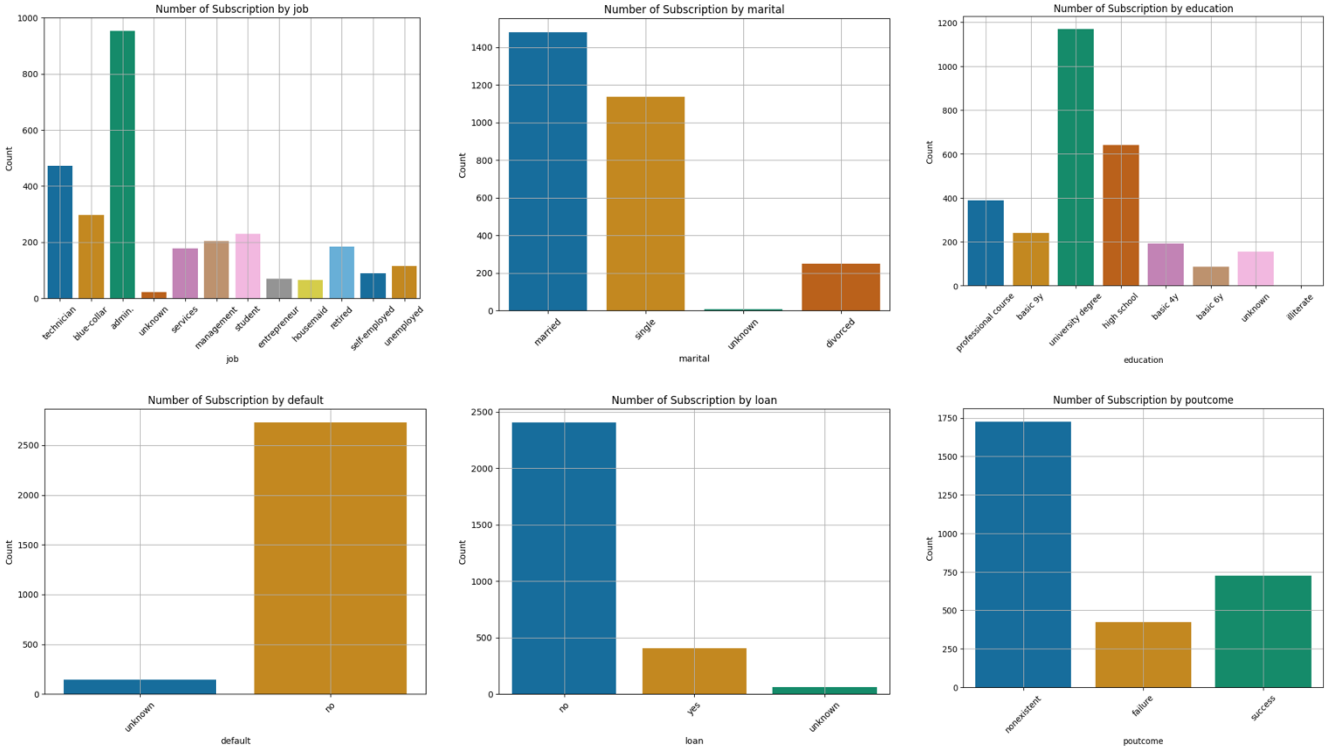
	housing	loan
job	Chi2: 28.06, p: 0.174	Chi2: 27.50, p: 0.193
marital	Chi2: 11.41, p: 0.076	Chi2: 4.15, p: 0.657
education	Chi2: 26.23, p: 0.024	Chi2: 13.98, p: 0.451
default	Chi2: 10.01, p: 0.040	Chi2: 3.91, p: 0.418
housing	-	Chi2: 35599.54, p: 0.000
loan	Chi2: 35599.54, p: 0.000	-
contact	Chi2: 260.38, p: 0.000	Chi2: 27.47, p: 0.000
poutcome	Chi2: 18.45, p: 0.001	Chi2: 1.05, p: 0.901

	contact	poutcome
job	Chi2: 537.26, p: 0.000	Chi2: 621.41, p: 0.000
marital	Chi2: 226.20, p: 0.000	Chi2: 134.93, p: 0.000
education	Chi2: 602.12, p: 0.000	Chi2: 168.17, p: 0.000
default	Chi2: 626.48, p: 0.000	Chi2: 439.18, p: 0.000
housing	Chi2: 260.38, p: 0.000	Chi2: 18.45, p: 0.001
loan	Chi2: 27.47, p: 0.000	Chi2: 1.05, p: 0.901
contact	-	Chi2: 2128.55, p: 0.000
poutcome	Chi2: 2128.55, p: 0.000	-

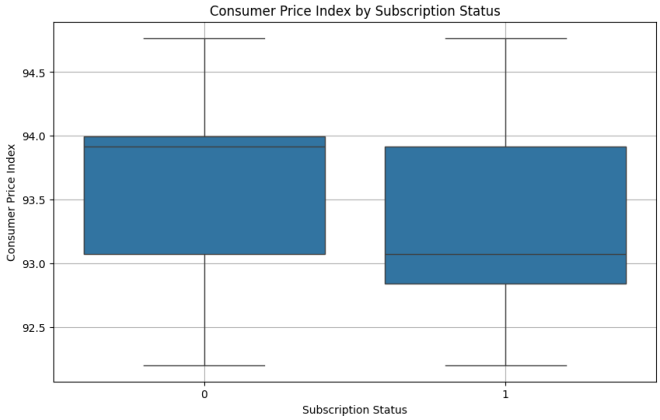
2. Subscription by age

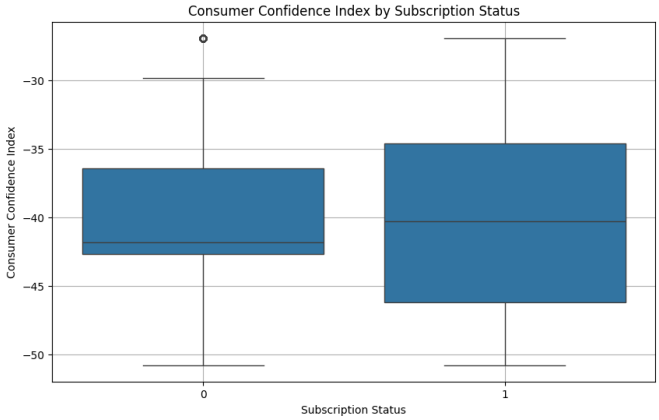


3. Categorical variables by subscription status

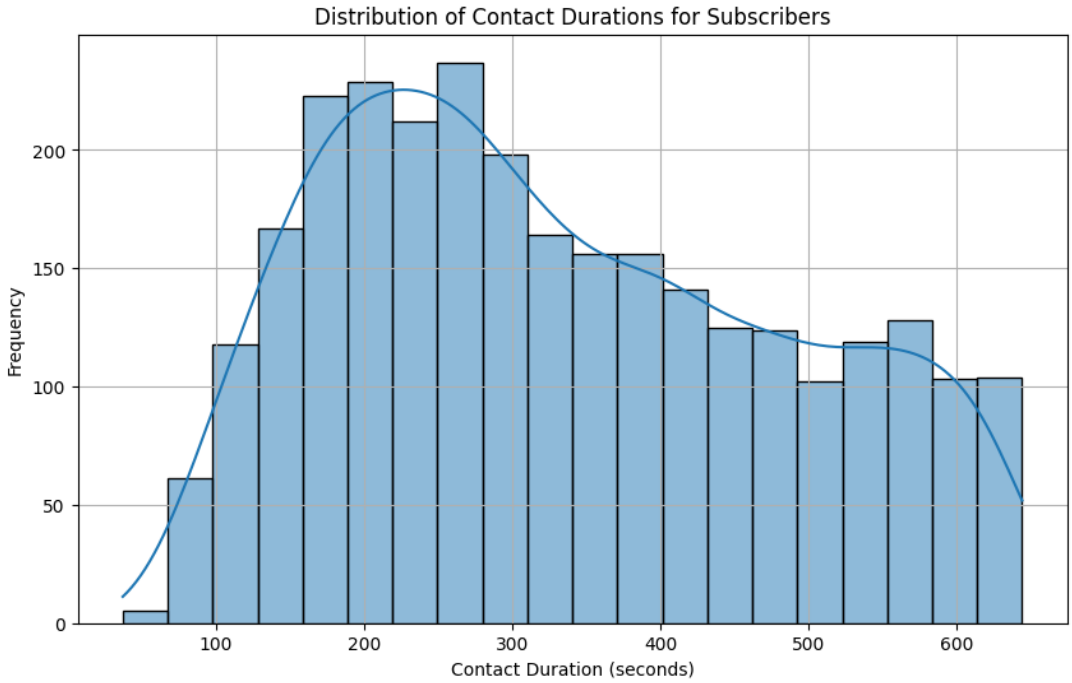


4. Subscription status by consumer price index and consumer confidence index





5. Distribution of contact duration by subscribers



6. Number of subscribers by month and days of week