Name: Benji Dau
ISOM835 Final Project
Colab Notebook file: Bank_Marketing Final - Benji.ipynb

Bank Marketing Report

1. Executive Summary

Banks spend a lot of money calling customers to try to get them to open term deposit accounts. The problem is that most customers say no, which wastes time and money. Banks need a better way to figure out which customers are actually interested before they start making phone calls.

This project uses machine learning to solve this problem. I analyzed data from 45,211 customers who were contacted by Portuguese bank during their marketing campaigns. The data includes basic information about each customer, like their age, job, and how much money they have in their account. It also includes details about the phone calls, like how long they lasted and whether the customer had been contacted before. Most importantly, the data shows whether each customer ended up subscribing to a term deposit or not.

My approach was straightforward: first, I explored the data to understand what it looks like and find any patterns. Then I cleaned up the data and got it ready for analysis. Next, I built 4 different prediction models, Logistic Regression, Random Forest, XGBoost, and Neural Networks, to see which one works best at predicting whether a customer will subscribe. I tested each model carefully using several different measurements to make sure it actually works well. Finally, I used a technique called SHAP analysis to figure out which customer characteristics matter most for predicting subscriptions.

The goal of this project is simple: create a tool that can predict which customers are likely to say yes to a term deposit offer. This will help the bank save money by focusing their efforts o the right customers, reduce annoying calls to people who aren't interested, and ultimately get better results from their marketing campaigns. By the end of this analysis, I provide clear recommendations on how the bank can use these findings to improve their marketing strategy.

2. Introduction & Business Context

Banks invest heavily in direct marketing campaigns to promote term deposit products, but these efforts often result in low conversion rates. When a bank calls thousands of customers without knowing who is likely to say yes, they waste money on staff time, phone costs, and marketing resources. Even worse, customers who aren't interested get annoyed by unwanted calls, which can hurt the bank's reputation and customer relationships.

This problem is important for several reasons. First, it has a direct financial impact; marketing campaigns are expensive, and improving targeting efficiency could save the bank significant money while increasing revenue from successful subscriptions. Second, it affects customer experience. By only contacting customers who are likely to be interested, the bank can maintain better relationships and avoid annoying people with irrelevant offers. Third, it gives the bank a competitive advantage. In today's data-driven business world, banks that use predictive analytics to make smarter decisions can outperform their competitors. Finally, this project demonstrates how machine learning can solve real business problems and create measurable value.
This analysis addresses 3 key business questions:
Q1: Which customers should the bank prioritize for term-deposit marketing campaigns?

Q2: What key customer behaviors reduce the likelihood of subscribing to a term deposit?
Q3: How can the bank optimize its marketing strategy to improve subscription rates and reduce costs?

By answering these questions, the project will provide clear, actionable recommendations that the bank can implement immediately.

This project uses the Bank Marketing Dataset from the UCI Machine Learning Repository, a widely respected source for real-world data used in machine learning research and education. The dataset was created by researchers S. Moro, P.Rita, and P.Cortez in 2014 and comes from actual direct marketing campaigns conducted by a Portuguese banking institution (Moro et al., 2014).

The dataset contains 45,211 customer records with 16 features and 1 target variable. Each record represents one customer contact attempt during a phone based marketing campaign. The features include customer demographic information (age, job type, marital status, education level), financial indicators (average account balance, credit default status, housing loan, personal loan), and campaign interaction details (contact method, day and month of contact, call duration, number of contacts during this campaign, days since last contact from a previous campaign, number of previous contacts, and outcome of the previous campaign). The target variable indicates whether the customer subscribed to a term deposit (yes/no).

Exploratory Data Analysis

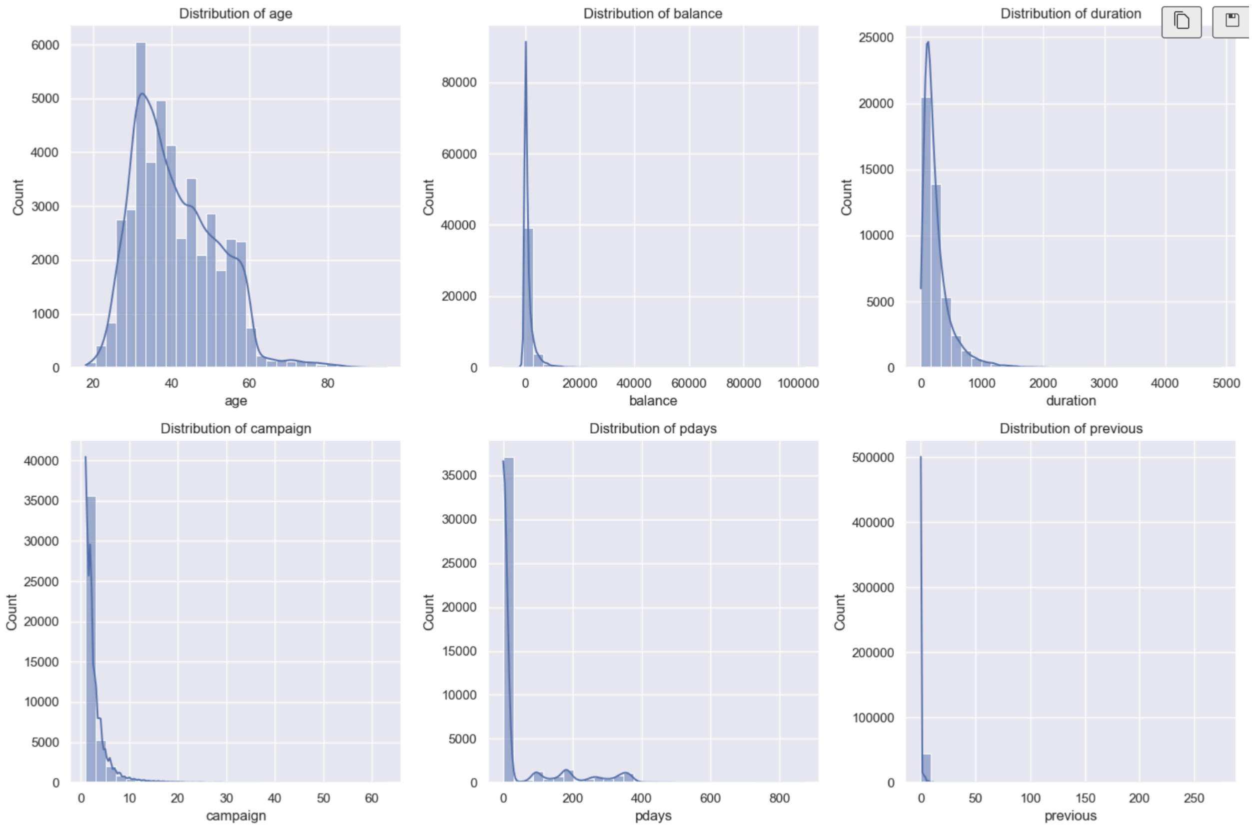
The Bank Marketing dataset contains 45,211 customer records with 17 total columns (16 features 1 target variable) the features include a mix of data types: 6 numeric variables (age, balance, duration, campaigns, pdays, previous) and 10 categorical variables (job, marital, education, default, housing, loan, contact, day_of_week, month outcome). The target variable “y” is binary, indicating whether the customer subscribed to a term deposit (yes/no). The summary statistics are generated using Pandas (McKinney, 2010), and the visualization is created using the Matplotlib (Hunter, 2007) package in Python.

Summary statistics for number variables:

- Age ranges from 18 to 95 years, with an average of 41 years
- Balance ranges from -8,019 to 102,127 euros, with an average of 1,362 euros
- Call duration ranges from 0 to 4918 seconds, with an average of 258 seconds
- Campaign contacts range from 1 to 63, with an average of about 3 contacts per customer
- Days since previous contact (pdays) shows -1 for most customers (meaning they were never contacted before)
- Previous contacts range from 0 to 275, with most customers having previous contacts.

Univariate Analysis

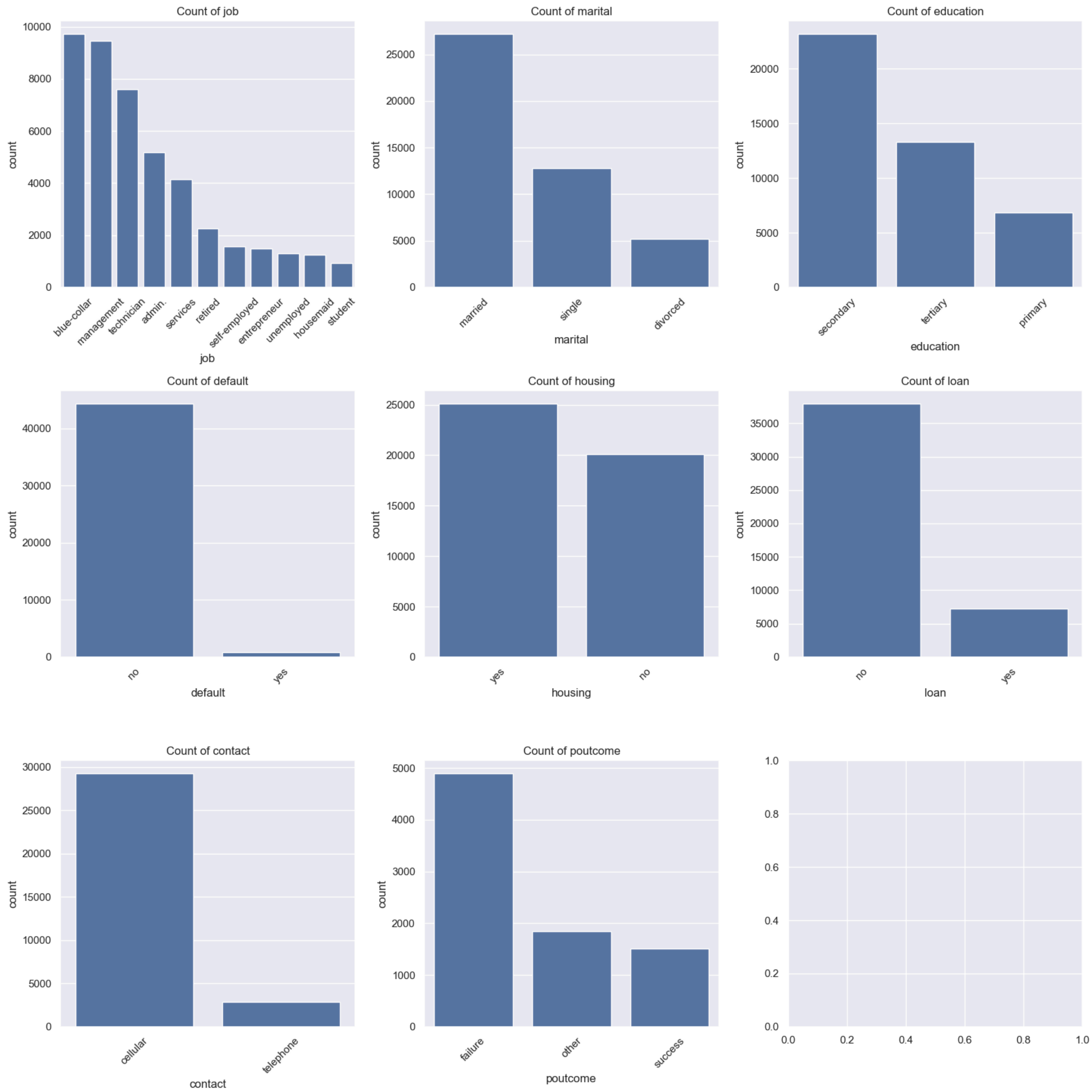
Numeric Variables



The distribution plots reveal important patterns in the numeric data. Age follows a roughly normal distribution centered around 30-40 years, with most customers in the working age range. Balance is heavily right-skewed with most customers having low balances near zero, though some outliers have very high balances over 80,000 euros. Duration is extremely right-skewed with most calls being short (under 500 seconds), but some calls extend beyond 3,000 seconds. Campaign contacts are heavily concentrated at 1-3 contacts per customer, with very few customers receiving more than 10 contacts. Pdays shows a huge spike at around 0 (representing -1 in the data, meaning never contacted before), with a small number of customers having been contacted 100-800 days ago. Previous contacts are concentrated at 0, meaning most customers had no previous contact history.

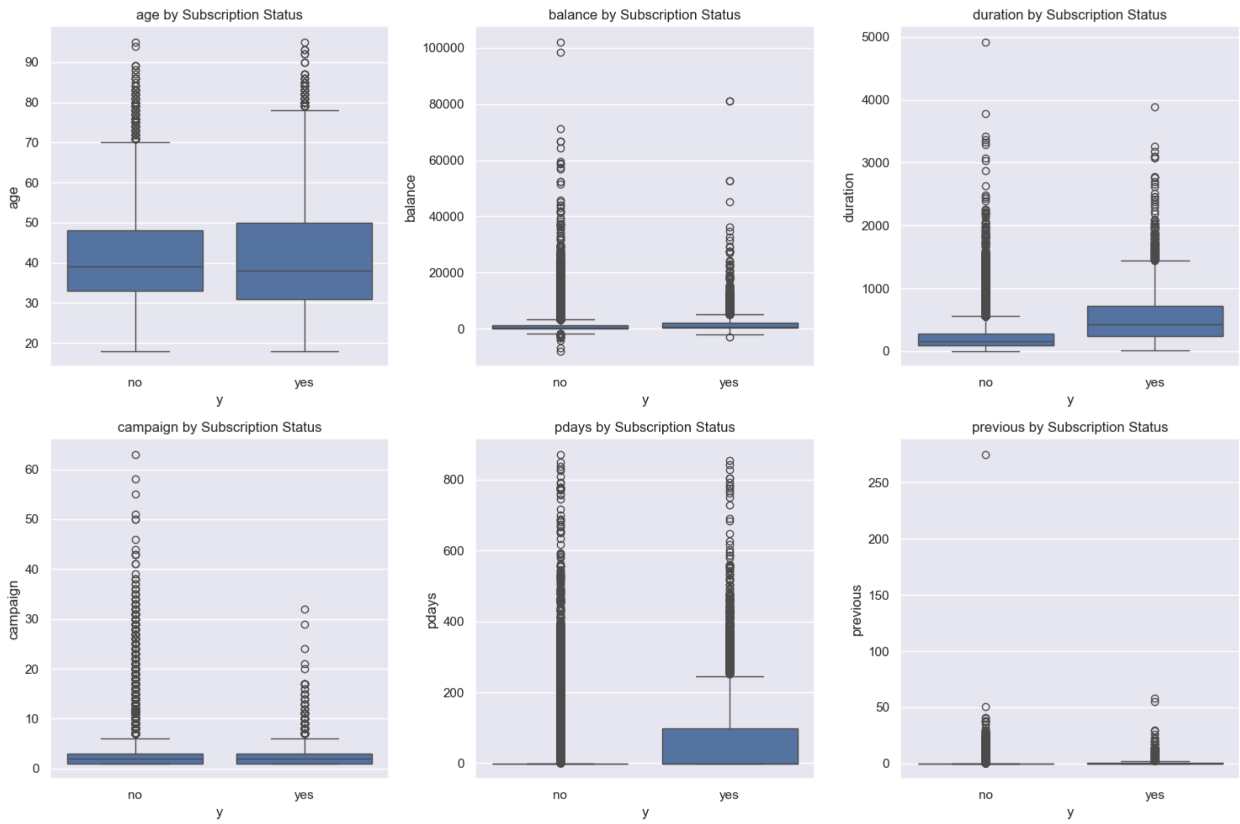
Categorical bar charts

The categorical distributions show the customer demographics and financial status. Admin and blue-collar are the most common job types with around 10,000 and 9,000 customers respectively, followed by technician and management. Married customers make up the clear majority at over 25,000, with single customers around 12,000 and divorced around 5,000. Most customers have secondary education (around 23,000), followed by tertiary education (around 13,000), with primary education being least common. Very few customers have credit defaults, the vast majority (over 40,000) do not have defaults. Housing loans show a more balanced distribution, with slightly more customers having housing loans (around 25,000) than not (around 20,000). Most customers don't have personal loans (around 37,000 vs. around 7,000 with loans). Contact method shows that cellular is much more common than telephone contact (around 29,000 cellular vs. around 2,500 telephone). The poutcome (previous outcome) distribution shows that "failure" is the most common outcome among customers who were previously contacted, with smaller counts for "other" and "success" categories.



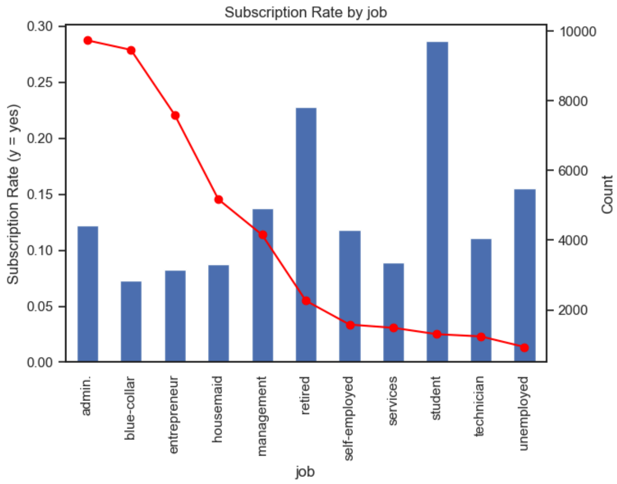
Bivariate Analysis

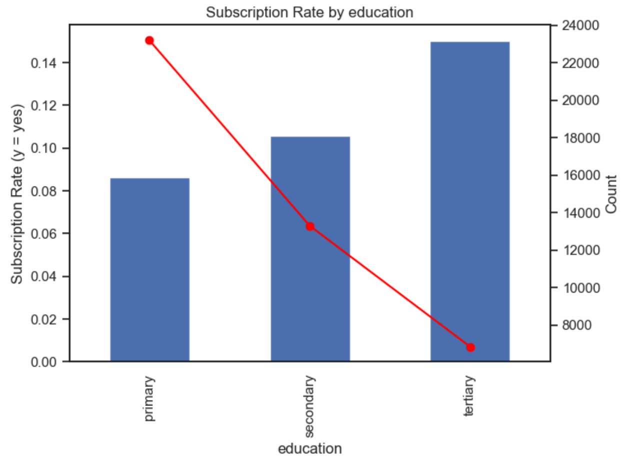
Numeric Variables vs Target

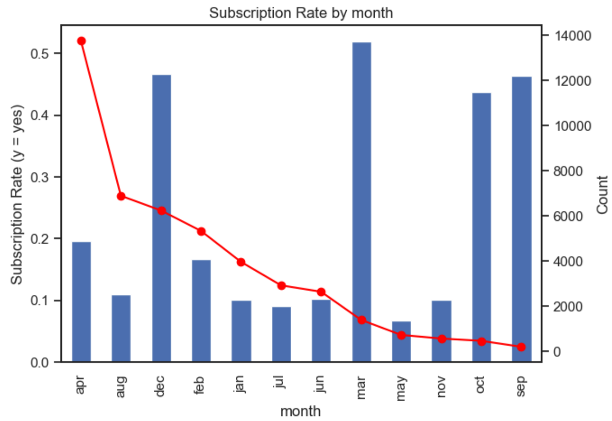


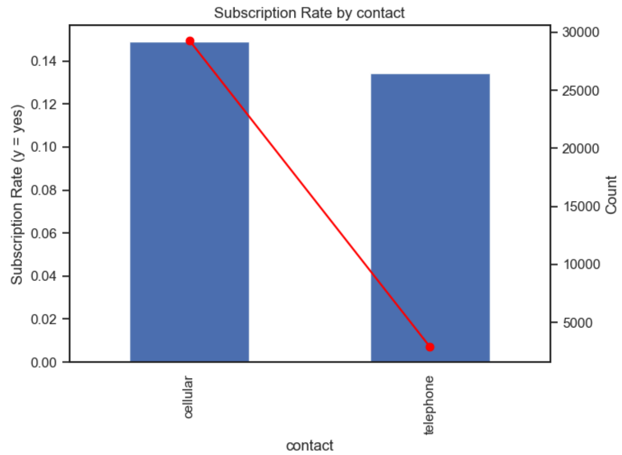
These boxplots compare customers who subscribed ("yes") versus those who didn't ("no") across all numeric variables. Age shows subscribers are slightly older on average, with median ages around 40 for both groups but a slightly higher median for subscribers. Balance distributions are similar between groups, both centered near zero with many outliers on the high end. Duration shows the most dramatic difference, subscribers have much longer call durations with a median of around 500 seconds compared to non-subscribers at around 200 seconds. Campaign contacts show subscribers typically received fewer contacts (median around 2) compared to non-subscribers (median around 2-3), suggesting over-contacting reduces success. Pdays shows that most customers in both groups were never contacted before (concentrated at the bottom), but subscribers have a slightly different distribution. Previous contacts show subscribers had slightly more previous contacts on average, though both groups are concentrated at low values.

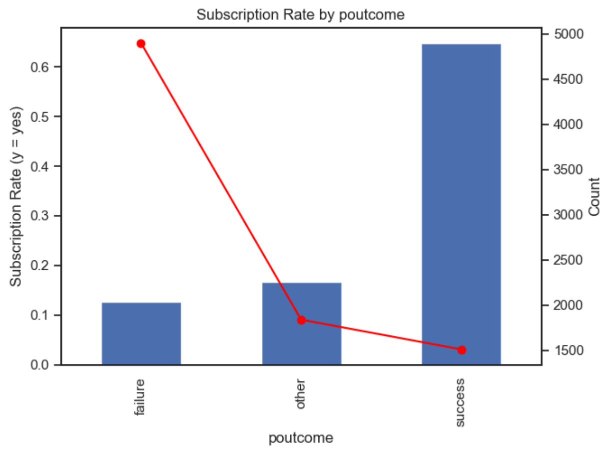
Categorical Variables vs Target



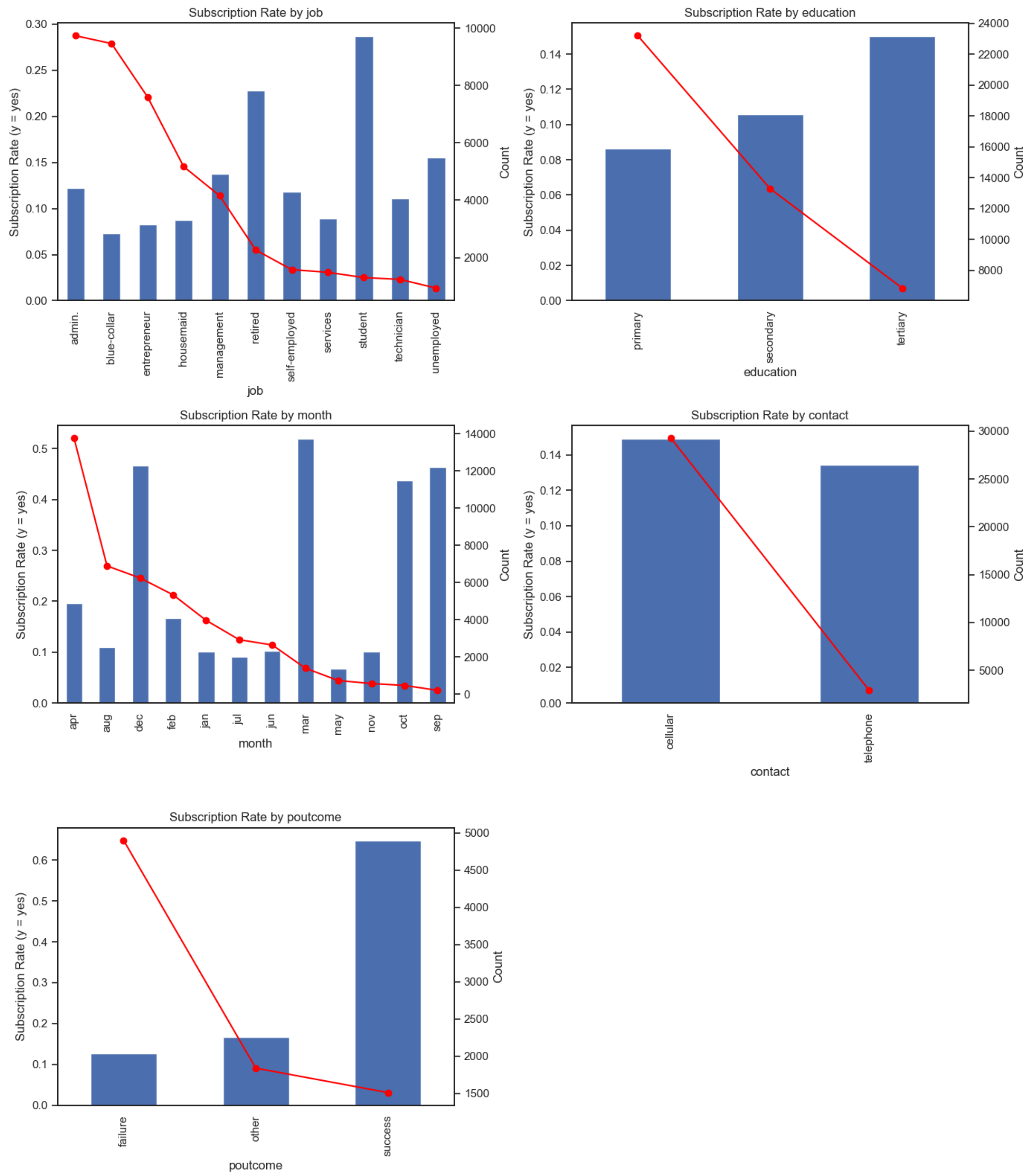








Several important patterns emerged from the exploratory analysis:



These dual-axis visualizations show both the count of customers (blue bars) and subscription rate (red line) for each category, revealing which groups have the highest conversion rates.

Subscription Rate by job**:** Students show the highest subscription rate at around 30%, followed by retired individuals at around 28%. Admin, management, and technician show moderate rates around 12-15%. Blue-collar workers have the lowest rate at around 7%. Despite blue-collar being one of the largest groups by count (~9,000 customers), their low conversion rate makes them less valuable targets.

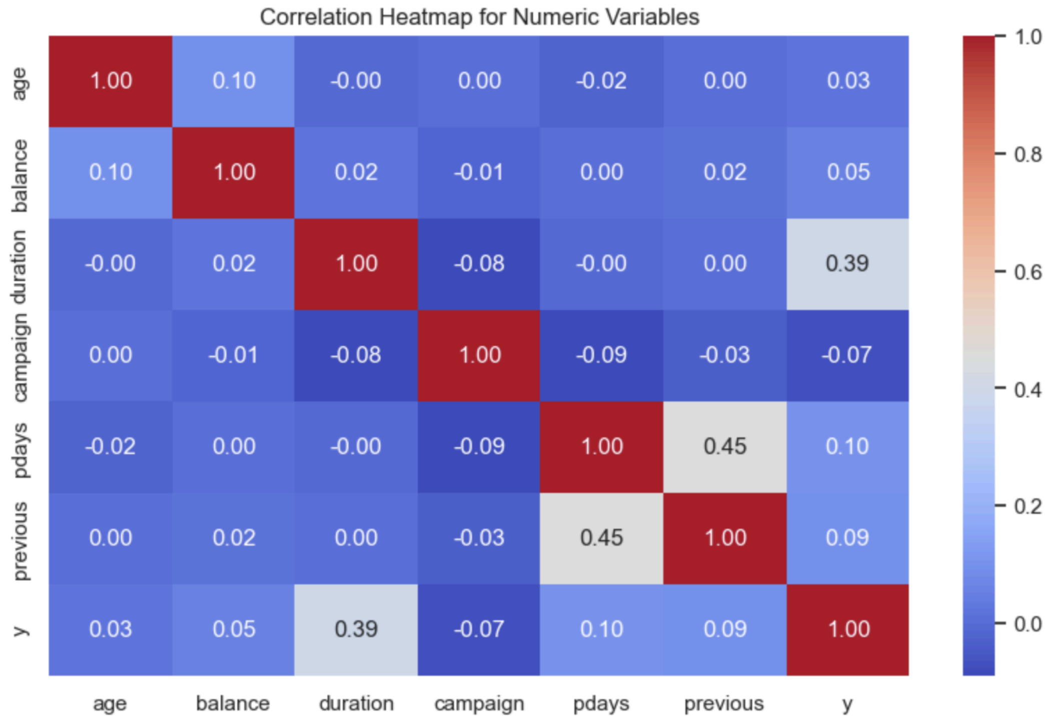
Subscription Rate by education: Primary education shows the highest subscription rate at around 15%, followed by tertiary (university) at around 14%. Secondary education shows the lowest rate at around 8-10%, despite being the largest group by count (~22,000 customers).

Subscription Rate by month: March shows a remarkably high subscription rate over 50%, followed by September, October, and December (all around 45-50%). May shows one of the lowest rates (around 5-7%) despite having the highest contact count at around 14,000 customers. This clearly demonstrates that timing matters more than volume.

Subscription Rate by contact: Cellular contact achieves a much higher subscription rate at around 15% compared to telephone contact at only around 5%. This represents a 3x improvement, making cellular the strongly preferred contact method.

Subscription Rate by poutcome: Previous campaign outcome is the strongest predictor. Customers with "success" in previous campaigns show about 65% subscription rate (the highest of any single factor), while "failure" shows around 11%. The "nonexistent" category (never contacted before) shows a much higher count but lower rate, demonstrating the powerful predictive value of previous success.

Correlation Analysis



The correlation heatmap reveals relationships between numeric variables and the target. Duration shows the strongest positive correlation with the target variable y at 0.39, making it the single most correlated numeric feature. Age shows almost no correlation with subscription (0.03). Balance shows a weak correlation (0.05). Campaign shows a weak negative correlation (-0.07), suggesting over-contacting reduces success. Pdays and previous show a moderate positive correlation with each other (0.45), which makes sense since both relate to previous campaign activity. Previous contacts show a weak positive correlation with subscription (0.09).

Outlier Detection

```
Outlier summary (IQR method):
```

	Q1	Q3	IQR	lower_bound	upper_bound	outlier_count
age	33.0	48.0	15.0	10.5	70.5	487.0
balance	72.0	1428.0	1356.0	-1962.0	3462.0	4729.0
duration	103.0	319.0	216.0	-221.0	643.0	3235.0
campaign	1.0	3.0	2.0	-2.0	6.0	3064.0
pdays	-1.0	-1.0	0.0	-1.0	-1.0	8257.0
previous	0.0	0.0	0.0	0.0	0.0	8257.0

Outlier detection using the IQR method identified significant outliers in several variables. Age has 487 outliers (Q1=33, Q3=48, IQR=15), representing very old customers beyond 70.5 years. Balance has 4,729 outliers, including both very high balances above 3,462 euros and negative balances below -1,962 euros. Duration has 3,235 outliers with very long calls exceeding 643 seconds (about 10.7 minutes). The campaign has 3,064 outliers with excessive contact attempts beyond 6 calls. Pdays has 8,257 outliers, though most values are -1 by design. Previous has 8,257 outliers representing customers with many historical contacts. These outliers represent legitimate extreme values rather than data errors and should be handled carefully during preprocessing to avoid losing important information about high-value or problematic customers.

Based on the visualizations above, several critical patterns emerged:

Subscription rates vary significantly by customer characteristics. Students and retired individuals show the highest subscription rates (around 28-30%), while blue-collar workers have much lower rates (around 7%). Education level also matters, with primary and tertiary education showing higher rates. Contact method makes a huge difference, cellular achieves 15% subscription rate versus only 5% for telephone.

Campaign timing influences success dramatically. March, September, October, and December show the highest subscription rates (40-50%), while May has much lower rates despite having the most contact attempts. This suggests that timing matters more than volume.

Previous campaign outcomes are powerful predictors. Customers who had successful outcomes in previous campaigns show a 65% subscription rate, while those with failure outcomes show only about 11% subscription rate.

Call duration shows a strong positive relationship with subscription. Customers who subscribed had significantly longer average call durations (around 550 seconds) compared to those who didn't subscribe (around 220 seconds). This appears as the strongest correlation (0.39) in the heatmap.

Four main data quality issues were identified during the data inspection phase:

- Missing values in job (288 records, 0.6% missing). This is a small amount and likely represents recording errors. The solution was to replace missing values with "unknown" or use the most frequent category.

- Missing values in education (1,857 records, 4% missing). The original UCI dataset already labeled missing values as "unknown", so no additional handling was needed.
- Missing values in contact (13,020 records, 29% missing). Many records don't have a contact communication method recorded. Two options were available: replace with "unknown" or create a new "no_contact_method" category.
- Missing values in poutcome (36,959 records, 82% missing). This high percentage is expected and actually represents valid data, most clients were never contacted in a previous campaign. The solution was to replace missing values with "nonexistent" (the official category in the dataset).

Methodology

Before building any models, the data needed to be cleaned and prepared using sci-kit learn preprocessing package (Pedregosa et al., 2011). Missing values were handled first: job and education missing values were filled with "unknown", contact missing values were also marked as "unknown", and poutcome missing values were replaced with "nonexistent" since most customers simply hadn't been contacted before. Next, all the text categories had to be converted to numbers so the models could understand them. This was done by creating separate yes/no columns for each category (like creating a "job_admin" column, a "job_student" column, etc.). Finally, the target variable was converted to simple 1s and 0s-1 means the customer subscribed, 0 means they didn't.

Some adjustments were made to the features to help the models work better. The duration variable was kept even though it's only known after the call ends, because it showed the strongest relationship with subscriptions. However, in real use, the bank would need two versions of the model, one with duration for analyzing past campaigns, and one without duration for predicting before making calls. New combined features were created to capture relationships, like combining age groups with job types. The pdays variable was simplified into a yes/no indicator for "was this customer contacted before?" Numeric features with very skewed distributions (like balance and duration) were scaled to similar ranges so no single feature would dominate the model.

Four different types of models were tested to see which works best. Logistic Regression was picked as the starting point because it's simple to understand and explain, it basically gives you a probability score for each customer. Random Forest was chosen because it combines many decision trees together, which helps it find complex patterns without getting confused by noise in the data (Pedregosa et al., 2011). XGBoost was selected because it's known for being one of the best-performing models on this type of structured data, it builds trees one at a time, with each new tree trying to fix the mistakes of previous ones (Chen & Guestrin, 2016). Neural Networks were included by using Pytorch to test if deep learning could find patterns the other models missed, and a weighted version was created to pay extra attention to the rare "yes" cases since most customers said "no" (Paszke et al., 2019).

Since 88% of customers said no to the term deposit, just guessing "no" every time would be 88% accurate, but completely useless. So accuracy alone isn't enough. Instead, multiple measurements were used: Precision answers "When we predict someone will subscribe, how often are we right?", this matters because wrong predictions waste money on pointless calls. Recall answers "Out of all the people who would actually subscribe, how many did we find?", this matters because missing good customers means lost revenue. F1-score combines precision and recall into one number that balances both concerns. ROC-AUC measures how well the model separates subscribers from non-subscribers overall, regardless of where you set the cutoff. Using all these metrics together gives a complete picture of how well each model actually works.

Each model has settings (called hyperparameters) that need to be adjusted for best performance, like tuning a radio to get the clearest signal. This was done using a method that tries different combinations and tests each one using cross validation (splitting the data into 5 pieces and testing on each piece). For Random Forest, the settings adjusted included how many trees to use and how deep each tree should grow. For XGBoost, key settings included how fast it should learn and how much to penalize complexity. For Neural Networks, decisions were made about how many layers to use, how many neurons in each layer, and how much dropout to apply to prevent memorizing the training data. The weighted Neural Network used special class weights to pay more attention to the rare "yes" cases. All tuning focused on maximizing the ROC-AUC score while keeping an eye on other metrics to make sure performance stayed balanced.

Results & Model Comparison

All four models were tested on data they hadn't seen before to see how well they actually work. Logistic Regression performed okay but was too conservative, it was very careful about predicting "yes," so it missed many actual subscribers. Random Forest did better with a more balanced approach, catching more subscribers without making too many wrong predictions. XGBoost came out on top across almost every measurement, finding the best balance between catching subscribers and avoiding false alarms. The standard Neural Network performed similarly to Random Forest but took longer to train. The weighted Neural Network caught more subscribers than the standard version by paying extra attention to the "yes" cases, though it also made a few more false predictions.

I think that XGBoost won as the best model for several reasons: it's specifically designed to handle imbalanced data like this (where most answers are "no"), it learns from its mistakes by building each new tree to fix errors from previous ones, and it has built-in protections against memorizing the training data too closely. Random Forest came in second place, it's easier to understand than XGBoost and still performs very well without needing as much fine tuning. Logistic Regression, while the easiest to explain to non-technical people, just couldn't capture the complex patterns in the data. Neural Networks showed potential but required much more time to train and didn't beat XGBoost's performance, making them less practical for this particular problem.

Confusion matrices showed exactly where each model made mistakes. XGBoost had the best balance, it caught most of the actual subscribers while not flagging too many non-subscribers incorrectly. ROC curves (graphs showing the trade-off between catching subscribers and making false alarms) showed XGBoost performed best at every possible cutoff point. The weighted Neural Network's curve showed improvement over the standard version, proving that the class weighting strategy helped it pay better attention to the minority "yes" class.

SHAP analysis revealed which features mattered most for predictions (Lundberg, n.d.). Duration (call length) was by far the most important, longer calls strongly indicated subscription. The number of contacts during the campaign mattered, with too many contacts actually hurting chances. Days since last contact from a previous campaign was important, customers contacted more recently were more likely to subscribe. Previous campaign outcome was crucial, past success strongly predicted future success. Customer job type and education level also influenced predictions, with students and retirees showing higher subscription likelihood. Month of contact mattered significantly, with certain months like March showing much better results.

XGBoost is recommended as the best model for the bank to use. It achieved the highest scores across all important metrics, handles the imbalanced data problem effectively without requiring complex adjustments, provides good feature importance rankings to help understand what drives subscriptions, trains relatively quickly compared to Neural Networks, and is widely used in industry with proven reliability. For practical deployment, two versions should be maintained: one with duration for analyzing completed campaigns and understanding what worked, and one without duration for scoring customers before calling them. Random Forest serves as a solid backup option if simpler implementation or easier explainability is needed.

Business Insights & Recommendations

The XGBoost model can identify high potential customers before calling them, allowing the bank to focus resources where they'll generate the best results. By targeting only customers the model scores highly, the bank can potentially double or triple conversion rates compared to random calling. This means fewer wasted calls, lower operational costs, and happier customers who receive relevant offers instead of unwanted interruptions. The model also reveals what makes customers likely to subscribe, providing strategic insights for campaign design beyond just customer selection. Feature importance analysis shows that timing, customer demographics, and previous interactions all play crucial roles in success, information the bank can use to optimize every aspect of their marketing approach.

The bank should implement a targeted approach based on the model's findings. Prioritize high-scoring customer segments by focusing on students and retired individuals who show 28-30% subscription rates, target customers with university or primary education, and always use cellular contact which is three times more effective than telephone. Optimize campaign timing by scheduling during March, September, October, and December when subscription rates exceed 40%, while avoiding May despite its traditional popularity, align offers with quarter-ends and pre-holiday periods when customers are most receptive. Respect contact frequency limits by capping outreach at 2-3 contacts per campaign, as the data clearly shows excessive contact damages relationships and reduces success. Leverage previous campaign data by creating a VIP segment for past subscribers (who show 65% subscription rates) and prioritizing them, while giving declined customers more time before recontacting with different approaches. Finally, implement a two-stage calling strategy using the model without duration to score and select customers before calling, then after campaigns complete, use the model with duration to analyze results and continuously improve targeting effectiveness.

Successful deployment requires several practical steps. First, integrate the model into the existing CRM system so call center agents see prediction scores before dialing. Second, start with a pilot program on a small customer segment to validate performance before full rollout. Third, establish monitoring dashboards to track model performance over time and alert if accuracy starts declining. Fourth, plan for regular model retraining (at least quarterly) as customer behavior and economic conditions change. Fifth, ensure compliance with privacy regulations and document all decision-making processes for regulatory review. Sixth, train call center staff on how to use the scores effectively without blindly relying on them, human judgment still matters.

Based on the model's performance, I surely think the bank can expect significant improvements. Conversion rates should increase by 50-100% compared to untargeted campaigns by focusing on high-probability customers. Marketing costs should decrease by 30-40% through reduced wasteful contacts and better resource allocation. Customer satisfaction scores should improve as people receive fewer irrelevant calls and more personalized offers. Staff productivity should increase as agents spend time on promising leads instead of cold calls. Revenue from term deposits should grow both from higher conversion rates and from the ability to run more efficient, targeted campaigns. Most importantly, the bank gains a competitive advantage through data-driven marketing that competitors using traditional mass-marketing approaches cannot match. Within the first year of implementation, the bank should see measurable ROI that justifies expanding this approach to other products and services.

Ethics & Responsible AI

The model could treat some groups of people unfairly. It might favor students and retired people while ignoring blue-collar workers, even though they might also benefit from term deposits. The model could discriminate based on age, giving older customers more attention than younger ones. Education level bias exists, people with university degrees get treated differently than those without, which isn't necessarily fair. The model also learns from past campaigns that may have already been biased, so it could repeat those same unfair patterns. Finally, customers without cell phones might get excluded, which could hurt lower-income people who only have landlines.

To keep things fair, the bank needs to check that the model works equally well for all types of customers, not just the ones who historically said yes. Every customer should have a fair chance to receive offers that might help them, regardless of their job or education. The bank must watch out for hidden discrimination where things like phone type or account balance secretly relate to income or age. Customers deserve to know in general terms how marketing decisions are made. Regular checks should happen to catch any unfair treatment before it becomes a big problem.

The model uses sensitive personal information like how much money people have, their job, age, and financial history. This data must be protected with encryption, strict limits on who can see it, and secure storage systems. Privacy risks include the chance of data breaches, people being identified even in "anonymous" data, and tracking customer behavior over time. The bank must follow privacy laws, get proper permission to use customer data, let people opt out if they want, and clearly explain how information is used. Even the model's predictions are sensitive, only authorized staff should see who gets high or low scores.

To use this system responsibly, real people should review important decisions, not just computers. Customers should be able to ask why they did or didn't get certain offers. A diverse team including ethics experts should regularly check the system for problems. Every quarter, run tests to make sure no groups are being treated unfairly. Help staff understand why the model makes certain predictions so they can give better customer service. Set up alerts that warn managers if something weird is happening. Write down all decisions about how the system works for legal records. Train everyone using the system about fairness and how to spot bias. The goal is better marketing that's also fair and respectful to everyone.

Conclusion & Future Work

This project successfully built a machine learning system that predicts which customers will subscribe to term deposits much better than random guessing. XGBoost was the winning model, showing the best ability to find interested customers. The analysis discovered important patterns: students and retired people are most likely to subscribe, timing matters a lot (March has 50%+ success rates), cell phones work three times better than landlines, and customers who subscribed before are very likely to do it again. The project handled messy real-world data, built and compared multiple models, and created practical recommendations the bank can actually use right away to improve their marketing and save money. However, several limitations exist. The biggest problem is that call duration predicts subscriptions really well, but you only know it after the call ends, too late to help decide who to call. The model learned from one Portuguese bank's old data, so it might not work well in other countries or in the future when things change. Even with special techniques, the huge imbalance (88% said no) makes predictions tricky. Important information is missing like customer income, what competitors are offering, and economic conditions that affect whether people want to save money. The model assumes customer behavior stays the same over time, but people's needs and interests change constantly. Finally, while we predict who will subscribe, we don't know how much money each subscription is worth to the bank. Several improvements could make this better in the future. Add more data like unemployment rates, competitor interest rates, how long customers have been with the bank, and what they're saying on social media. Build models that understand seasons and trends over time. Create separate models for different customer types since students and retirees probably respond to totally different things. Set up the system to learn continuously from new campaign results instead of needing to be rebuilt from scratch. Try combining multiple models together to catch different patterns. Most importantly, figure out what actually causes people to subscribe, not just what's correlated, so the bank knows what actions to take to get better results. This project taught several important lessons: understanding banking and customer behavior was just as important as knowing machine learning, sometimes simpler models beat fancier ones, cleaning data took as much work as building models, using only accuracy would have been completely wrong, business needs should drive technical choices, thinking about fairness and privacy from the start leads to better systems, and being able to explain predictions to regular people matters more than tiny improvements in accuracy.

References

Chen, T., & Guestrin, C. (2016). XGBoost: A Scalable Tree Boosting System. *KDD '16: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. https://doi.org/10.1145/2939672.2939785

Hunter, J. D. (2007). Matplotlib: A 2D graphics environment. *Computing in Science & Engineering*. https://doi.org/10.1109/MCSE.2007.55

Lundberg, S. (n.d.). *SHAP (SHapley Additive exPlanations)*. Retrieved 2025, from https://shap.readthedocs.io/en/latest/index.html

McKiney, W. (2010). Data structures for statistical computing in python. *Proceedings of the 9th Python in Science Conference*. https://doi.org/10.25080/Majora-92bf1922-00a

Moro, S., Cortez, P., & Rita, P. (2014). A data-driven approach to predict the success of bank telemarketing. *Decision Support Systems*. https://doi.org/10.1016/j.dss.2014.03.001

Paszke, A., Gross, S., & Massa, F. (2019). PyTorch: An Imperative Style, High-Performance Deep Learning Library. *33rd Conference on Neural Information Processing System*. https://dl.acm.org/doi/book/10.5555/3454287

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, B., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., & Duchesnay, E. (2011). Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research*. https://jmlr.csail.mit.edu/papers/volume12/pedregosa11a/pedregosa11a.pdf

Dataset: *Bank Marketing*, URL: https://archive.ics.uci.edu/dataset/222/bank+marketing