

Samantha,

Imagine I went grocery shopping for you without any context. I wouldn't know what to get, would I? However, if I peeked into your fridge and pantry, I'd have a better idea of what you might need. By reviewing your previous shopping lists, the task becomes even simpler. This analogy parallels our approach to direct marketing using data-driven strategies. While relying on gut instinct or previous campaign frameworks has its merits, leveraging a data-driven approach allows us to target customers with precision akin to knowing exactly what's on your shopping list. The bank faces several challenges in attracting customers for its time deposits. Directed marketing efforts, while focused, often provoke privacy concerns, potentially creating negative attitudes toward the bank. This sensitivity is compounded by an overarching challenge: the average success rate of our campaigns hovers around only 8%, amidst escalating competition. We can implement a lot of different approaches to attract new customers and strengthen relationships with existing ones. Regular customer feedback sessions, such as surveys and interviews, help the bank align with customer needs and unveil new opportunities for product enhancements. We could participate in community events and form strategic partnerships with local businesses, which could enhance our presence but also attract customers who value community-focused businesses, thereby expanding its customer base through shared values and benefits. While these strategies offer considerable potential, their success hinges on our ability to effectively track and analyze their impact. Which brings us to the data we have at hand right now. I analyzed detailed records from 45,211 calls made during previous marketing campaigns from 2008 to 2010. This dataset is really comprehensive—it covers everything from the customer's job and age to their financial status and how they've responded to our previous campaigns. Each record includes outcomes like whether the customer signed up for our services after the call, giving us clear indications of what's worked in the past. Let us assume the cost to contact a customer is \$2 by phone. If we make even 2% off a deposit and a customer deposits \$1000 for a year then the bank would make \$20 in profit. I started off with implementing Logistic Regression. Imagine we're trying to predict if a customer will buy a new phone based on things like their age, how much they earn, and what phone they currently use. This model helps us make this prediction by giving each factor a certain weight and it accurately predicted customer decisions about 90% of the time and suggested that job type and account balance are significant predictors. It does estimate a net profit of \$12,598 but it doesn't tell us why customers might say yes or no, which limits how personalized our approach can be. Next I decided to use Decision tree. it works by following a set of yes/no questions to reach a decision. It was pretty good at picking out actual subscribers and gave us clear rules, like

if a customer is from a certain job or has a certain balance, they might be more likely to subscribe. This model was less accurate overall but provided clear rules on why customers might subscribe. It achieved a balance of precision and recall with a net profit of \$8,468. The direct insights from Decision Trees make them valuable for planning specific conversations in targeted campaigns, even though they aren't as broadly accurate as other models.

If a Decision Tree is like asking one friend for advice, using a Random Forest is like asking a group of friends. This method combines the opinions of many decision trees to make a final decision. This is our best choice. It combined the clarity of Decision Trees with much better accuracy—about 90% correct in predictions. Plus, it was the best at using our budget wisely, suggesting who to call to likely increase our profits. It combines clarity with enhanced accuracy, correctly predicting outcomes about 90% of the time and was the most effective in terms of cost-benefit, suggesting a net profit of \$14,530. The Random Forest model, in particular, allows us to target our efforts more precisely, promising better response rates and higher profitability. We expect to see improvements in customer engagement and conversion rates, which translates to higher overall profitability. Data-driven approaches in marketing campaigns, as demonstrated through our model analysis, underscore the importance of precisely targeting and personalizing communications based on detailed customer insights. By employing models such as Logistic Regression, Decision Trees, and Random Forests, we learn to optimize resources, adapt strategies based on real-time data, and enhance overall campaign effectiveness. These models reveal which demographics and behaviors are most likely to respond, which contact strategies are most efficient, and how past interactions influence current campaign outcomes, ensuring that marketing efforts are not just broad but sharply focused on maximizing both customer engagement and profitability. For the upcoming cross-selling campaign, the target audience will include customers aged 50-65, particularly those in executive, managerial, or professional roles, indicating higher financial stability and investment potential. The focus will be on individuals with account balances in the top 25% of our database, and those who have previously engaged in lengthy discussions over financial services. This campaign will strategically time its outreach during periods identified as having higher engagement rates from past data, maximizing the probability of successful conversions and optimizing marketing resources effectively.

Appendix

Logistic Regression: I started with Logistic Regression because it's well-suited for binary outcomes, like predicting whether a customer will subscribe to a term deposit or not. It performed impressively, achieving a ROC AUC score of 0.906. This high score shows it can clearly distinguish between those likely to subscribe and those who are not. The model also maintained a solid overall accuracy of about 0.90. However, Logistic Regression showed some limitations in the context of a targeted marketing campaign. While it generated a respectable net profit of \$12,598, it wasn't the highest-earning model we tested. More importantly, despite its statistical accuracy, it didn't offer much insight into customer behavior, which is crucial for crafting targeted marketing strategies.

```
Cross-validated ROC AUC scores: [ 0.91302486 0.89943567 0.90469153 0.9045194 0.9128375 ]
Mean ROC AUC score: 0.9069017896233136
Mean accuracy: 0.9022894253945672 ± 0.0020553757120199555
Mean precision: 0.6499998722551081 ± 0.01858132300404023
Mean recall: 0.34326323854929336 ± 0.008876448392747002
Mean f1: 0.44918732874393197 ± 0.010429587092911329
Mean roc_auc: 0.9069017896233136 ± 0.005273113490088282
```

Figure 1: Performance Metrics for Logistic Regression

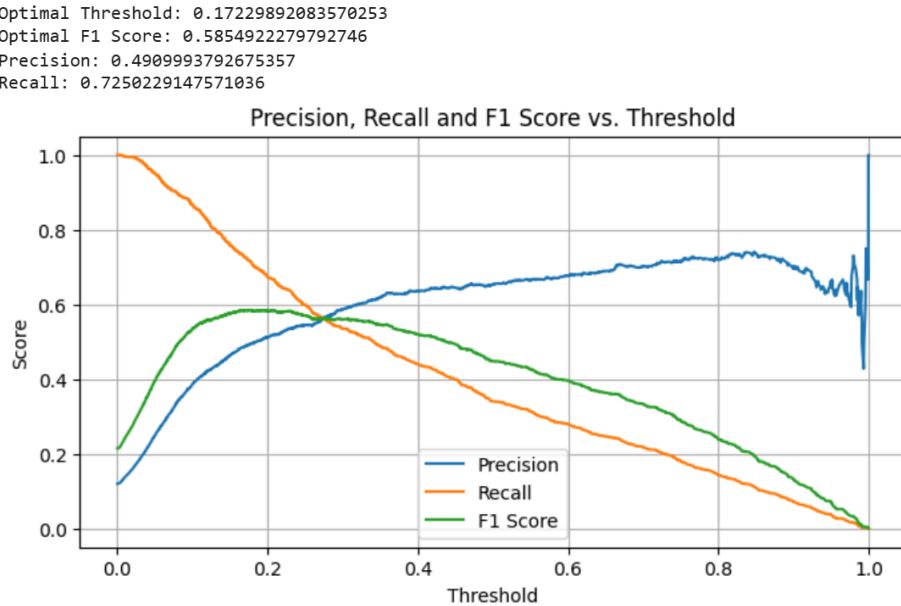


Figure 2: Precision, Recall, F1 score vs threshold for Logistic Regression

| | Feature | Coefficient |
|----|------------------|-------------|
| 47 | duration | 1.081595 |
| 27 | contact_unknown | -0.429839 |
| 42 | poutcome_success | 0.369145 |
| 25 | contact_cellular | 0.334243 |
| 48 | campaign | -0.273094 |
| 34 | month_jun | 0.262870 |
| 35 | month_mar | 0.195566 |
| 22 | housing_yes | -0.169417 |
| 21 | housing_no | 0.169417 |
| 33 | month_jul | -0.167386 |

Figure 3: Most significant features for Logistic Regression

Decision Tree: The Decision Tree model had a ROC AUC score of 0.703, which is moderate and shows less effectiveness in discrimination compared to Logistic Regression. However, it demonstrated a fair balance of precision and recall, doing quite well at identifying actual subscribers when they indeed subscribed, with precision around 0.476 and recall at 0.491. The transparency and straightforward rules of the Decision Tree make it a bit better suited for targeted campaigns than Logistic Regression. It allows marketers to understand exactly why a customer is predicted to subscribe, which is beneficial for tailored communication. Despite these advantages, the model's financial performance was less impressive, generating a lower net profit of \$8,468.

```

Cross-validated ROC AUC scores for Decision Tree: [0.71178672 0.69395993 0.71435442 0.68967723 0.70362682]
Mean ROC AUC score for Decision Tree: 0.702681021123033
Optimal Threshold for Decision Tree: 1.0
Optimal F1 Score for Decision Tree: 0.48353631032927374
Precision for Decision Tree: 0.47602131438721135
Recall for Decision Tree: 0.4912923923006416

```

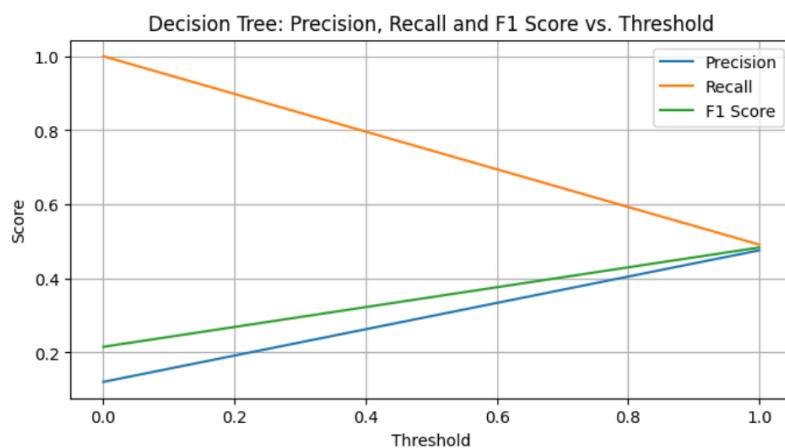


Figure 4: Performance Metrics and Precision, Recall, F1 score vs threshold for Decision Tree

| | Feature | Importance |
|----|------------------|------------|
| 47 | duration | 0.268723 |
| 45 | balance | 0.109701 |
| 42 | poutcome_success | 0.091093 |
| 44 | age | 0.089177 |
| 46 | day | 0.083872 |
| 49 | pdays | 0.046619 |
| 48 | campaign | 0.031183 |
| 21 | housing_no | 0.017261 |
| 50 | previous | 0.015686 |
| 35 | month_mar | 0.013505 |

Figure 5: Most significant features for Decision Tree

Logistic Regression Cost-Benefit Analysis:
Net Profit: \$ 12598
Total Contacts: 1611
Successful Conversions: 791

Decision Tree Cost-Benefit Analysis:
Net Profit: \$ 8468
Total Contacts: 1126
Successful Conversions: 536

Figure 6: Cost Benefit Analysis for Logistic Regression and Decision Tree

Random Forest: The Random Forest model stood out in every key metric. It notched the highest ROC AUC score at 0.928, indicating superior capability to differentiate between potential subscribers and non-subscribers. It also led in the financial analysis, earning the highest net profit of \$14,530. Moreover, the model reached an impressive accuracy of approximately 0.903 and maintained a balanced F1 score of 0.508, demonstrating strong precision and recall.

I chose the Random Forest as my final model for two main reasons: it delivers the highest financial returns and combines the interpretability of a Decision Tree with much better accuracy.

Cross-validated ROC AUC scores for Random Forest: [0.92978657 0.92611228 0.92722176 0.92887896 0.92804963]
Mean ROC AUC score for Random Forest: 0.9280098418692477

Figure 7: Performance Metrics for Random Forest

```
Precision: 0.6583941605839416
Recall: 0.41338221814848763
F1 Score: 0.5078828828828829
Accuracy: 0.9033506579674887
ROC AUC: 0.9249866176896393
```

Figure 8: Precision, Recall, F1 Score, and Final Accuracy for Random Forest

```
Random Forest Cost-Benefit Analysis:
Net Profit: $ 14530
Total Contacts: 2065
Successful Conversions: 933
```

Figure 9: Cost Benefit Analysis for Random Forest

| | Feature | Importance |
|----|---------------------|------------|
| 1 | num_balance | 0.702046 |
| 0 | num_age | 0.250029 |
| 3 | cat_job_blue-collar | 0.004577 |
| 10 | cat_job_student | 0.004184 |
| 7 | cat_job_retired | 0.004173 |
| 15 | cat_marital_married | 0.003890 |
| 6 | cat_job_management | 0.003866 |
| 16 | cat_marital_single | 0.003805 |
| 11 | cat_job_technician | 0.003698 |
| 2 | cat_job_admin. | 0.003284 |

Figure 10: Most significant features for Random Forest

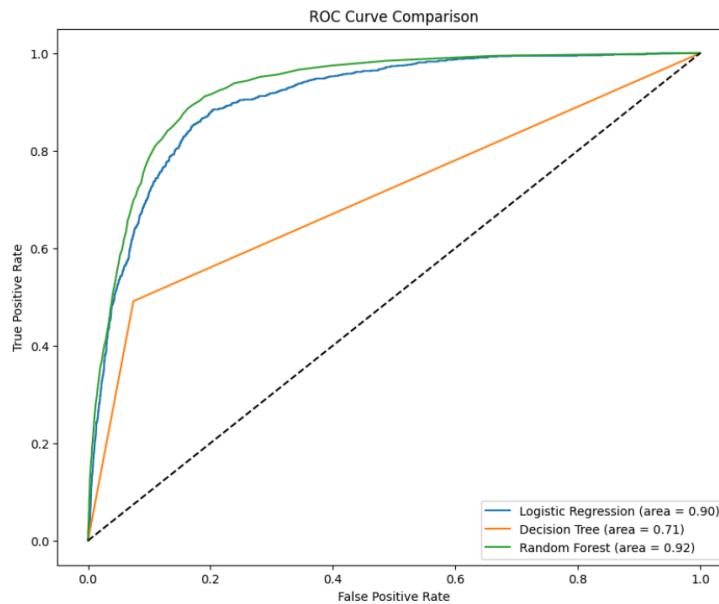


Figure 11: ROC Curve Comparison for all three models

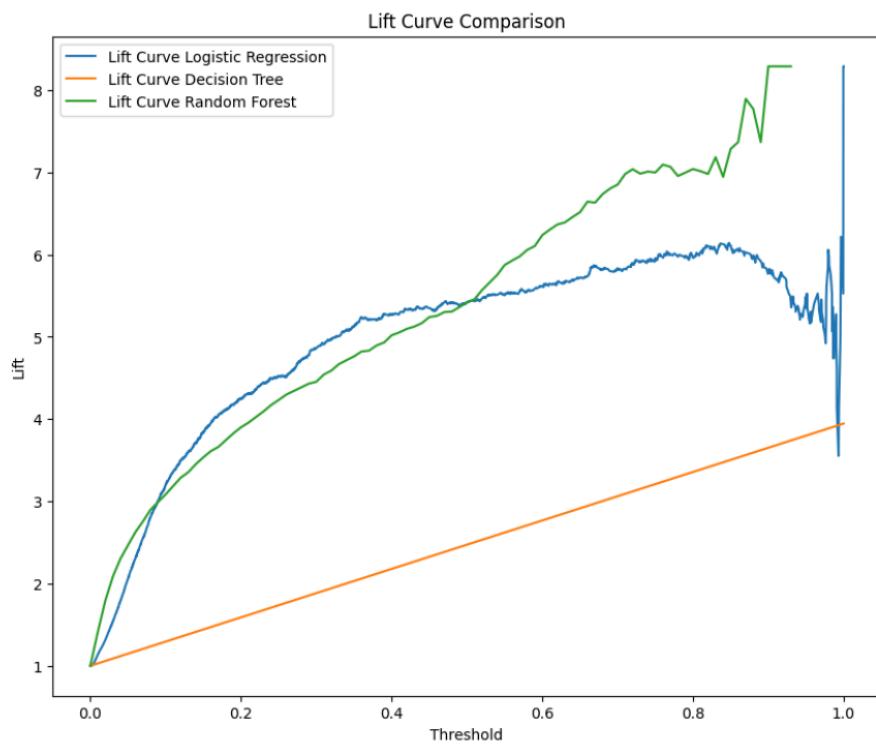


Figure 12: Lift Curve Comparison for all three models

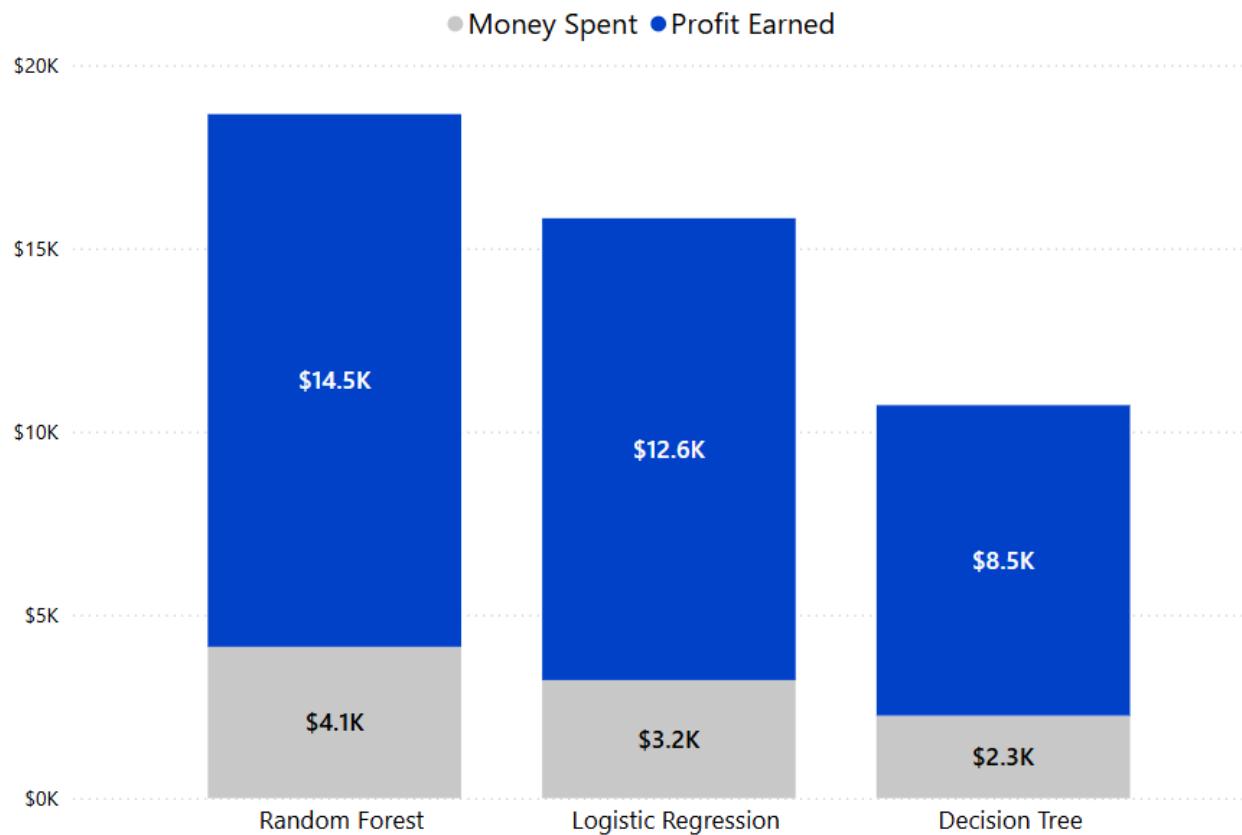


Figure 13: Cost Benefit Comparison for all models

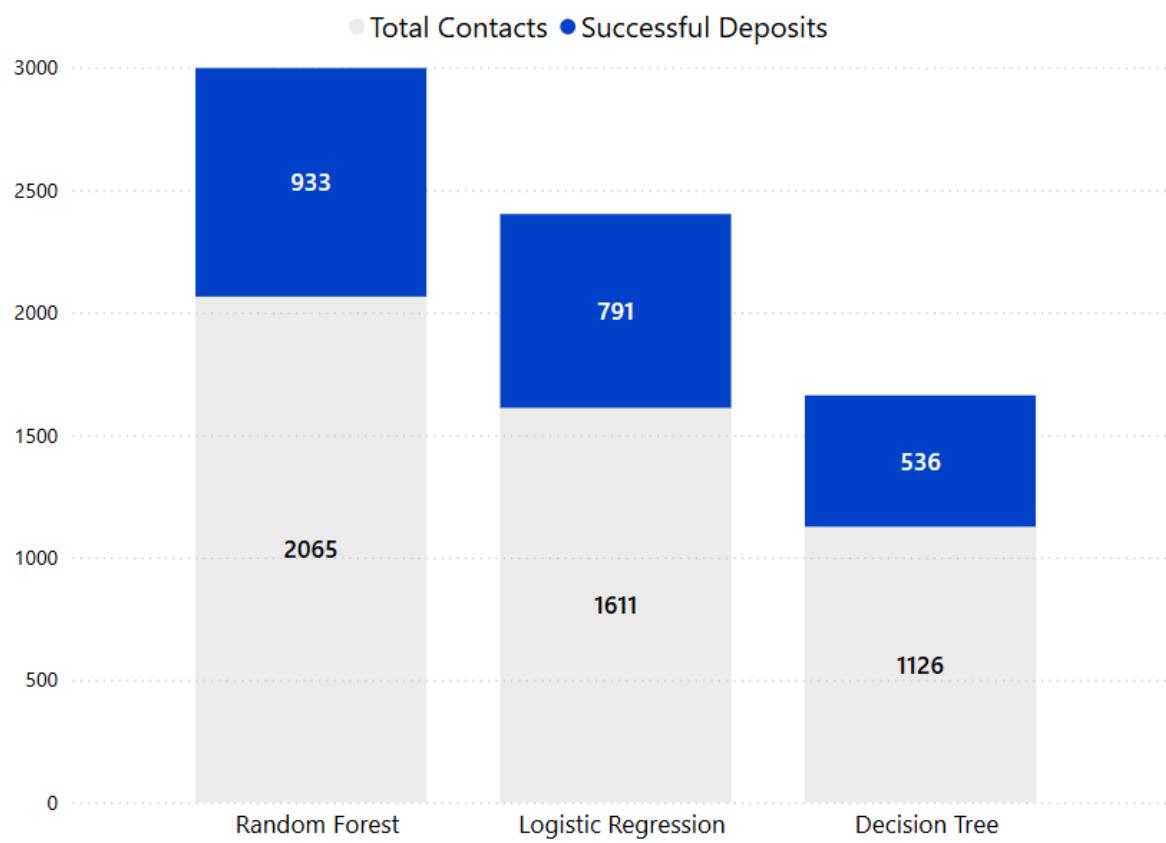


Figure 14: Total Contacts vs. Successful Deposits comparison for all models