Data Analysis Project For Marketing Campaign

Assessment Task 3

Project Report

Fatemeh Elyasifar 25589351
Statistical Thinking for Data Science
TD School
University of Technology Sydney

Table of Contents

Summary and Introduction

Summary

In today's competitive telecommunication industry, understanding customer behaviour and predicting their response to marketing campaigns is critical for success. This project focuses on predicting the likelihood of customers subscribing to a new telecommunication plan.

To tackle this, statistical learning models were developed to predict the likelihood of a campaign's success for individual customers. The models analyse customer data, uncovering key patterns that allow for targeted marketing approaches.

The models achieved a prediction accuracy of over 90%. Additionally, three significant insights were identified to guide future marketing strategies, improving customer targeting and campaign effectiveness.

These results provide clear, data-driven strategies for optimising marketing efforts, helping the company make more informed decisions and maximise campaign success.

Introduction

Problem Statement

The company faces difficulty identifying customer segments that are most responsive to marketing campaigns, leading to inefficient strategies and missed opportunities. The central problem addressed in this analysis is the prediction of customer subscription to a new plan based on historical data.

Rationale

Accurately predicting customer behaviour is essential for optimising marketing efforts and enhancing overall campaign success. Understanding customer preferences can significantly improve targeting, engagement, and return on investment.

Project Aims and Objectives

The main aim of this project is to build statistical models that predict the success of marketing campaigns for customers. The specific objectives of research are as follows:

1. Identifying the customer segments most responsive to marketing campaigns
2. Deriving effective business strategies

To address these objectives, it was hypothesised that economic indicator, contact methods, and campaign engagement, would significantly influence subscription likelihood. The goal was to develop predictive models to classify customers by their likelihood of subscribing.

Methodology

Methods Overview

This project utilised four machine learning models to predict customer subscription likelihood:

1. Logistic Regression
2. Decision Tree
3. Multi-layer Perceptron (MLP) Classifier
4. XGBoost

Additionally, Maximum Likelihood Estimation (MLE) was used to extract meaningful business insights from the historical data.

Methods Details

Data Acquisition

The dataset includes numerical and categorical data (Appendix A, Figure 1), with the target variable “y” as a Boolean datatype. It was stored in CSV format and loaded as a data frame.

Data Preprocessing

Data preprocessing was crucial, addressing missing values and balancing the dataset using SMOTE due to the imbalanced target variable. SMOTE generated new minority class samples, enhancing the model's performance on imbalanced data.

```
Number of rows before SMOTE: 41168

Number of rows after SMOTE: 66660

Increase in number of rows after SMOTE: 25492
```

Figure 1. Number of Rows Before & After SMOTE

Modelling and Evaluation

Several machine learning models were tested, including Logistic Regression, Decision Tree, MLP Classifier, and XGBoost, each selected for its strengths in classification tasks. Logistic Regression was chosen for its interpretability, Decision Tree for capturing non-linear patterns, MLP Classifier for complex mappings, and XGBoost for high performance and handling imbalanced data. **Principal Component Analysis (PCA)** was not effective, as it did not reduce features significantly and decreased accuracy (Appendix B, Figure 1). All models were evaluated on Accuracy, Precision, Recall, and F1-Score, and showed good generalisation, with low standard deviations in cross-validation scores indicating no signs of overfitting.
Additionally, a Generalised Linear Model (GLM) using Maximum Likelihood Estimation (MLE) was built to identify key predictors of subscription likelihood. The performance metrics (Figure 2) were consistent across folds, suggesting stable model performance and good generalization.

Validation Scores:

	Fold	Validation Score
0	1	0.061127
1	2	0.066389
2	3	0.065842
3	4	0.066358
4	5	0.062873

Figure 2. Validation Scores

```
Cross-Validation Scores: [0.88608663 0.8823364  0.88683668 0.88823254 0.88138772]
Mean CV Score: 0.8849759913792081
Std CV Score: 0.0026511859250314975
Training Accuracy: 0.8849197419741974
0.8793879387938794
[[6492  762]
 [ 846 5232]]
              precision    recall  f1-score   support

           0       0.88      0.89      0.89      7254
           1       0.87      0.86      0.87      6078

    accuracy                           0.88     13332
   macro avg       0.88      0.88      0.88     13332
weighted avg       0.88      0.88      0.88     13332
```

Figure 3. Logistic Regression Result

```
Cross-Validation Scores: [0.92105757 0.91936996 0.91965123 0.92086263 0.91683075]
Mean CV Score: 0.9195544289028612
Std CV Score: 0.0015121258744701105
Training Accuracy: 1.0
0.9234923492349235
[[6781  574]
 [ 446 5531]]
              precision    recall  f1-score   support

           0       0.94      0.92      0.93      7355
           1       0.91      0.93      0.92      5977

    accuracy                           0.92     13332
   macro avg       0.92      0.92      0.92     13332
weighted avg       0.92      0.92      0.92     13332
```

Figure 4. Decision Tree Result

```
Cross-Validation Scores: [0.91693231 0.91318207 0.91421339 0.91092358 0.91064229]
Mean CV Score: 0.9131787280305582
Std CV Score: 0.002309939868552465
Training Accuracy: 0.9231735673567357
[[6574  697]
 [ 392 5669]]
              precision    recall  f1-score   support

           0       0.94      0.90      0.92      7271
           1       0.89      0.94      0.91      6061

    accuracy                           0.92     13332
   macro avg       0.92      0.92      0.92     13332
weighted avg       0.92      0.92      0.92     13332
```

Figure 5. MLP Result

```
Cross-Validation Scores: [0.93905869 0.93315207 0.94215263 0.93886545 0.93933427]
Mean CV Score: 0.9385126232836809
Std CV Score: 0.0029354036219477572
Training Accuracy: 0.9574707470747075
[[6774  459]
 [ 341 5758]]
              precision    recall  f1-score   support

           0       0.95      0.94      0.94      7233
           1       0.93      0.94      0.94      6099

    accuracy                           0.94     13332
   macro avg       0.94      0.94      0.94     13332
weighted avg       0.94      0.94      0.94     13332
```

Figure 6. XGBoost Result

Results

Key Findings

- Consumer price index and duration of contact showed a strong positive impact.
- Among non-parametric models, the **Decision Tree** performed the best.
- Among parametric models, the **MLP** performed the best.
- Overall, the **Decision Tree** was the most effective model for predicting subscription likelihood.

In-depth Results

Model Comparison

The performance metrics of each model are shown in the bar chart. Key findings include:

- **Logistic Regression** performed well but lagged more complex models, particularly in recall and F1-Score.
- **Decision Tree** emerged as the top-performing model, with an accuracy of 0.985, precision of 0.982, recall of 0.985, and an F1-Score of 0.983. Its high recall and F1-Score make it ideal for identifying likely subscribers.
- **MLP Classifier** performed slightly lower than Decision Tree and XGBoost.
- **XGBoost** showed strong performance with precision of 0.949 and recall of 0.960, making it effective at balancing false positives and negatives.

Right Model:
For maximising accuracy and identifying subscribers: **Decision Tree**
For balanced predictions: **XGBoost**
For simpler models: **Logistic Regression**
For robust but slightly lower performance: **MLP Classifier**

Given these results, the **Decision Tree** was selected as the optimal model due to its superior performance and effectiveness in identifying potential subscribers after applying SMOTE to balance the dataset.

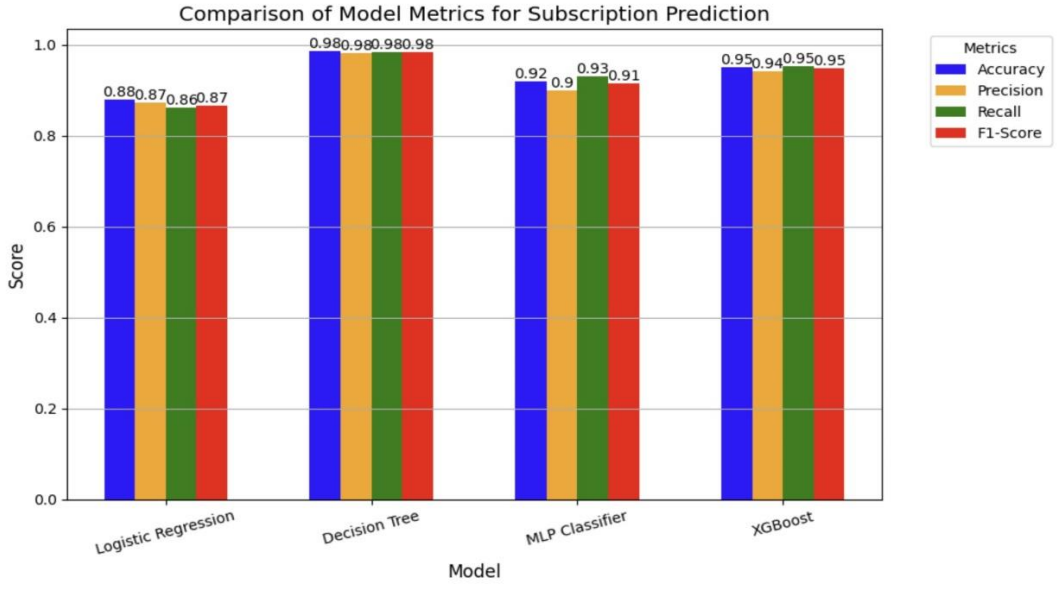

Figure 7. Comparison of Model Metrics

Business Insights and Recommendations

Based on the provided results of the GLM using MLE (Figure 4), four key insights emerged:

1. **Duration of Contact:** The coefficient of 0.0045 and high z-score (56.257), indicating it is a **highly significant predictor**, suggesting that **longer contact durations** are associated with a higher likelihood of a positive outcome.

2. **Contact Method:** A negative coefficient of -0.9716 with a low p-value (0.000) indicates that certain contact methods negatively affect the outcome, suggesting a need to optimise the contact strategy.

3. **Economic Indicators:**

    - Emp.var.rate has a **strong negative impact** (-0.9203), suggesting that a lower employment variation rate increases the likelihood of a positive outcome.
    - Cons.price.idx has a **positive impact** (1.1165), indicating that a higher consumer price index is linked with a better response.

    These findings highlight the importance of economic trends in influencing the predicted outcome.

4. **Previous Campaigns:** Multiple prior contacts negatively impact current campaign responses, possibly due to contact fatigue.

Also, the Pseudo R-squared (0.2385) indicates a moderate model fit.

Generalized Linear Model Regression Results

Dep. Variable:	y	No. Observations:	32935
Model:	GLM	Df Residuals:	32914
Model Family:	Binomial	Df Model:	20
Link Function:	Logit	Scale:	1.0000
Method:	IRLS	Log-Likelihood:	-7130.7
Date:	Sun, 10 Nov 2024	Deviance:	14261.
Time:	08:56:34	Pearson chi2:	2.91e+07
No. Iterations:	7	Pseudo R-squ. (CS):	0.2385
Covariance Type:	nonrobust		

	coef	std err	z	P>|z|	[0.025	0.975]
const	-78.7796	20.239	-3.892	0.000	-118.448	-39.111
age	0.0090	0.002	4.294	0.000	0.005	0.013
job	0.0071	0.006	1.119	0.263	-0.005	0.020
marital	0.1198	0.041	2.950	0.003	0.040	0.199
education	0.0737	0.016	4.566	0.000	0.042	0.105
default	0.3660	0.072	5.068	0.000	0.224	0.508
housing	-0.0241	0.045	-0.539	0.590	-0.112	0.064
loan	-0.1336	0.064	-2.093	0.036	-0.259	-0.008
contact	-0.9716	0.077	-12.566	0.000	-1.123	-0.820
month	-0.0560	0.015	-3.804	0.000	-0.085	-0.027
day_of_week	0.0134	0.016	0.853	0.394	-0.017	0.044
duration	0.0045	7.99e-05	56.257	0.000	0.004	0.005
campaign	-0.0272	0.012	-2.187	0.029	-0.052	-0.003
pdays	-0.0021	0.000	-15.717	0.000	-0.002	-0.002
previous	-0.1600	0.058	-2.744	0.006	-0.274	-0.046
poutcome	-0.2645	0.088	-3.000	0.003	-0.437	-0.092
emp.var.rate	-0.9203	0.080	-11.501	0.000	-1.077	-0.763
cons.price.idx	1.1165	0.126	8.882	0.000	0.870	1.363
cons.conf.idx	0.0483	0.007	6.872	0.000	0.035	0.062
euribor3m	0.2795	0.121	2.314	0.021	0.043	0.516
nr.employed	-0.0052	0.002	-2.665	0.008	-0.009	-0.001

Figure 8. Summary of GLM Result

Conclusion and References

Conclusion

This project demonstrated the effectiveness of machine learning models in predicting customer subscription likelihood, providing actionable insights for marketing strategies. Among the models, Decision Tree performed best, excelling in precision, recall, and F1-Score, making it ideal for this task. With the insights gained, the company can refine its marketing approach. These recommendations are expected to boost subscription rates, ROI, and customer loyalty.

Future work could focus on exploring additional features, refining models for specific customer segments, and integrating real-time data for dynamic predictions. Overall, this analysis supports data-driven decision-making in telecommunications, aligning marketing strategies with customer needs.

References

- Wickham, H. (2014). Tidy data. Journal of Statistical Software, Articles, 59(10), 1–23. https://doi.org/10.18637/jss.v059.i10
- Breiman, Leo.“Random Forests.” Machine Learning 45 (1). Springer: 5-32 (2001).
- Holte, Robert C. “Very simple classification rules perform well on most commonly used datasets.” Machine learning 11.1 (1993): 63-90.
- Ayman H. Abdel-aziem*, Tamer H. M. Soliman. A Multi-Layer Perceptron (MLP) Neural Networks for Stellar
- Classification: A Review of Methods and Results. https://www.researchgate.net/publication/373239041_A_Multi-Layer_Perceptron_MLP_Neural_Networks_for_Stellar_Classification_A_Review_of_Methods_and_Results
- Alice Dong. (2024). TeleCom_Data_1.csv [Dataset]. The real-world dataset provided by the professor.

Appendices

Appendix A: Overview of the dataset

	age	job	marital	education	default	housing	loan	contact	month	day_of_week	...	campaign	pdays	previous	poutcome	emp.var.rate	cons.price.idx	cons.conf.idx	euribor3m	nr.employed	y
0	40	admin.	married	basic.6y	no	no	no	telephone	may	mon	...	1	999	0	nonexistent	1.1	93.994	-36.4	4.857	5191.0	no
1	56	services	married	high.school	no	no	yes	telephone	may	mon	...	1	999	0	nonexistent	1.1	93.994	-36.4	4.857	5191.0	no
2	45	services	married	basic.9y	unknown	no	no	telephone	may	mon	...	1	999	0	nonexistent	1.1	93.994	-36.4	4.857	5191.0	no
3	59	admin.	married	professional.course	no	no	no	telephone	may	mon	...	1	999	0	nonexistent	1.1	93.994	-36.4	4.857	5191.0	no
4	41	blue-collar	married	unknown	unknown	no	no	telephone	may	mon	...	1	999	0	nonexistent	1.1	93.994	-36.4	4.857	5191.0	no
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
41175	29	unemployed	single	basic.4y	no	yes	no	cellular	nov	fri	...	1	9	1	success	-1.1	94.767	-50.8	1.028	4963.6	no
41176	73	retired	married	professional.course	no	yes	no	cellular	nov	fri	...	1	999	0	nonexistent	-1.1	94.767	-50.8	1.028	4963.6	yes
41177	46	blue-collar	married	professional.course	no	no	no	cellular	nov	fri	...	1	999	0	nonexistent	-1.1	94.767	-50.8	1.028	4963.6	no
41178	56	retired	married	university.degree	no	yes	no	cellular	nov	fri	...	2	999	0	nonexistent	-1.1	94.767	-50.8	1.028	4963.6	no
41179	44	technician	married	professional.course	no	no	no	cellular	nov	fri	...	1	999	0	nonexistent	-1.1	94.767	-50.8	1.028	4963.6	yes

41180 rows × 21 columns

Figure A1. The Dataset

Appendix B: Models accuracy considering PCA

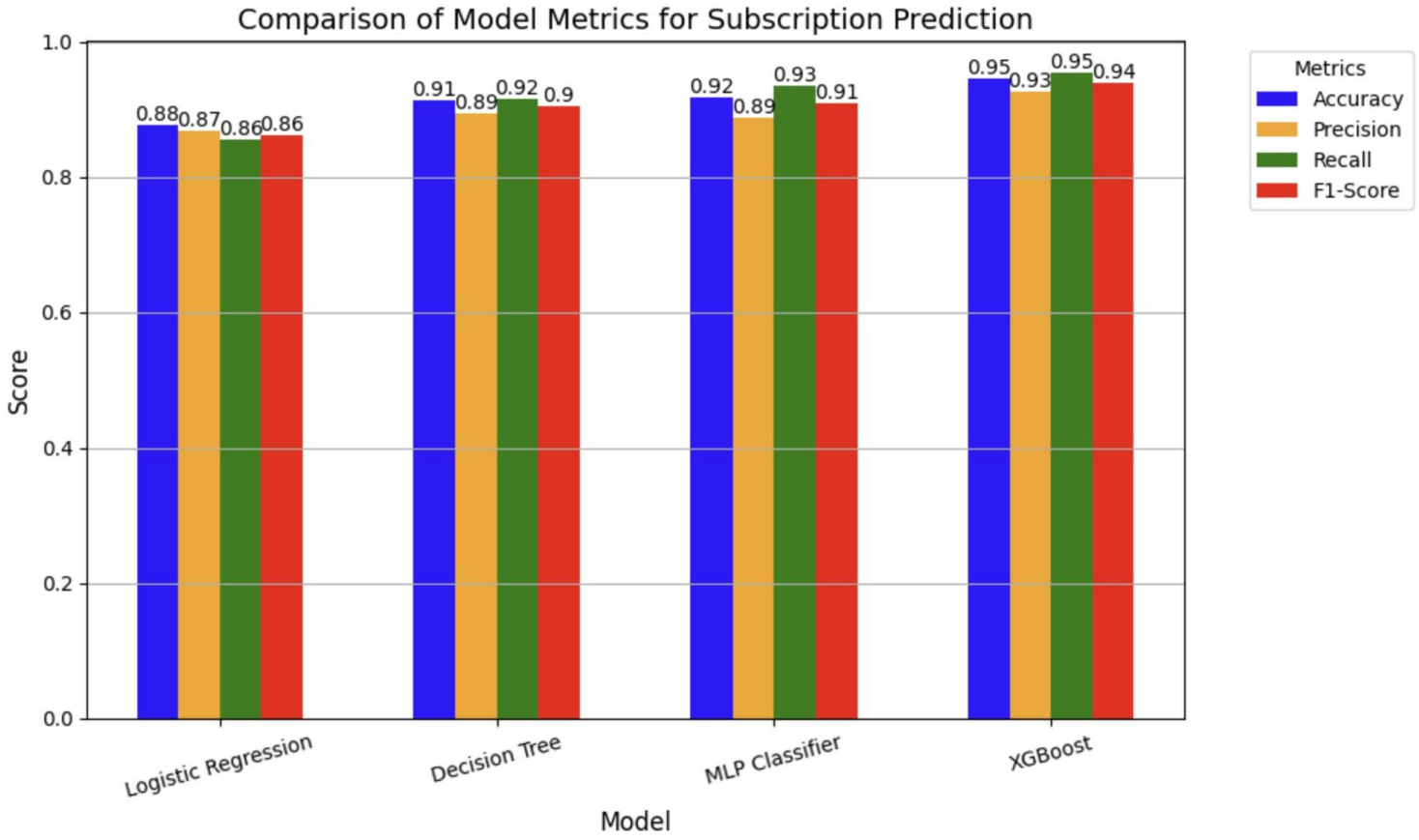

Figure B1. Models Accuracy Using PCA

Appendix C: Mean CV Score & Standard Deviation

	Model	Mean CV Score	Standard Deviation
0	Logistic Regression	0.884976	0.002565
1	Decision Tree	0.920398	0.003015
2	MLP Classifier	0.914848	0.003381
3	XGBoost	0.937406	0.002268

Figure C1. Validation Scores

Appendix D: ROC Curve Analysis for Model Performance

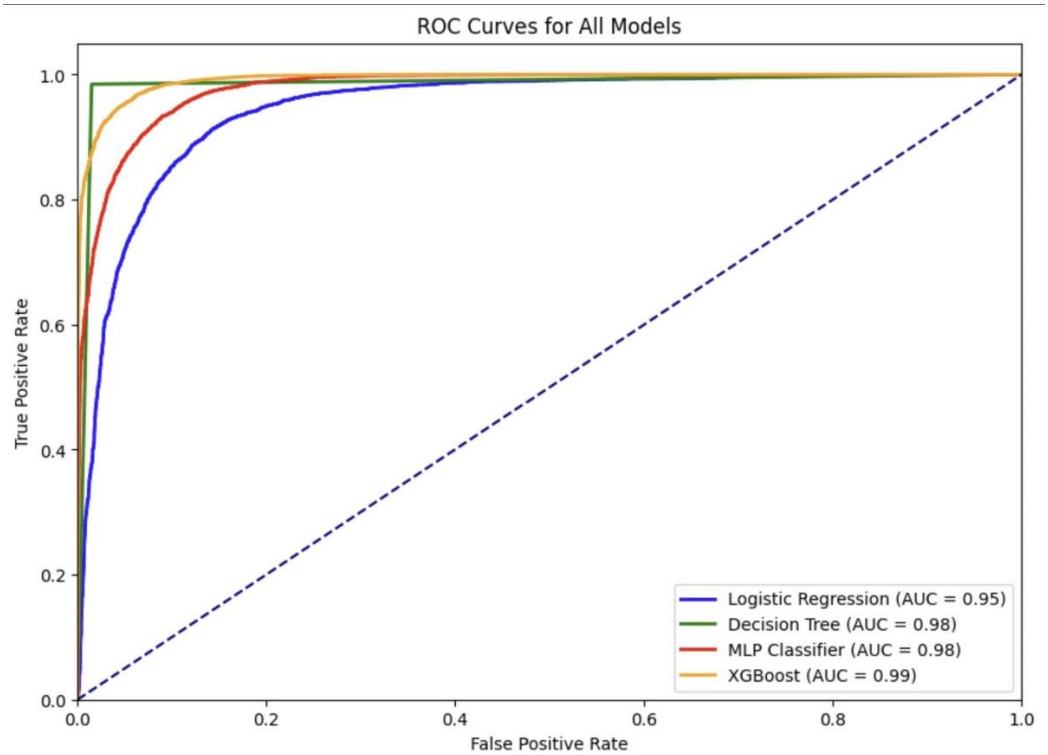

Figure D1. Evaluating Model Discrimination Ability