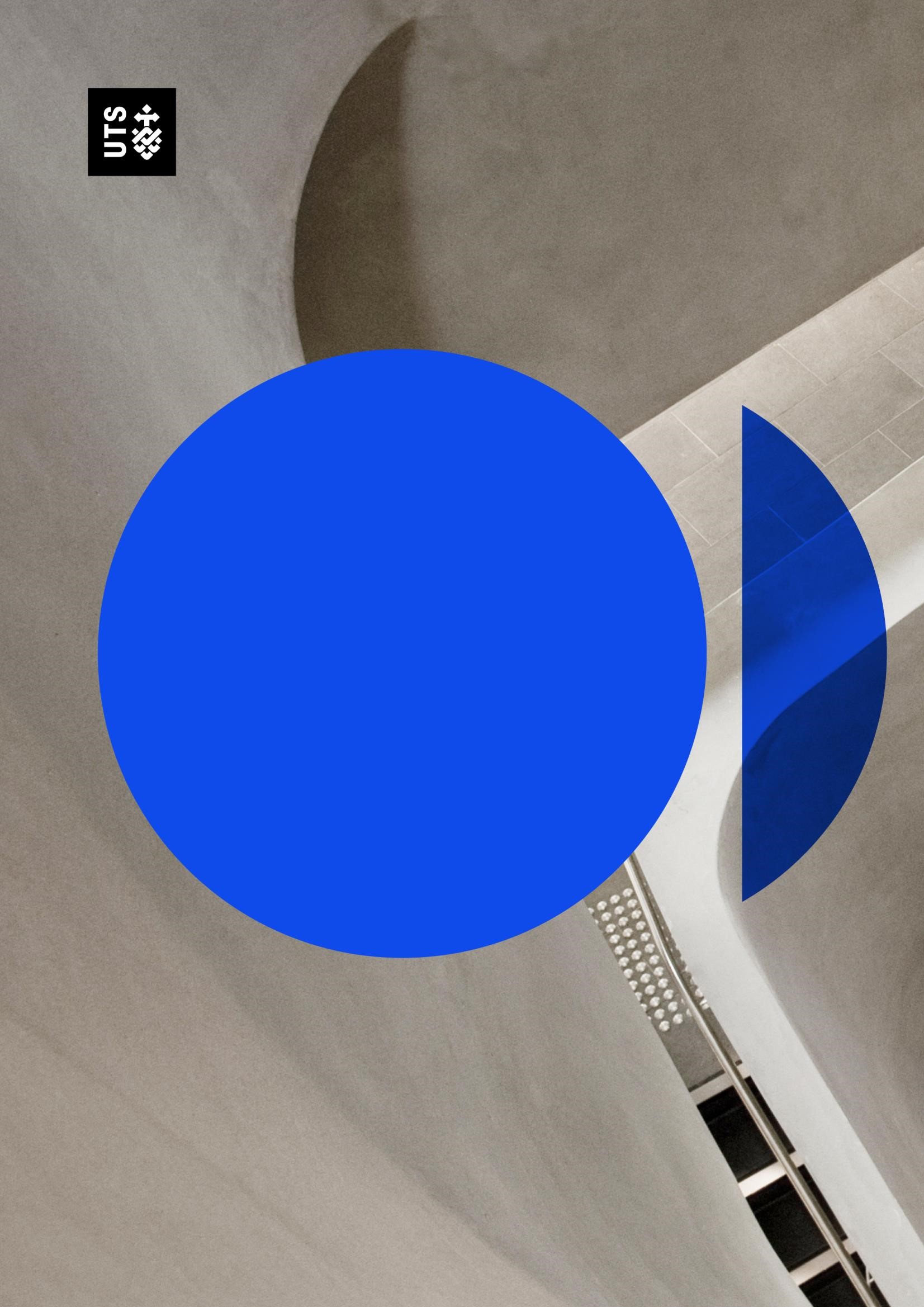
STDS 36103

–

AT2B2 project report structure example



UTS

CRICOS 00099F

Like and Subscribe; Statistical Approaches to Predicting Phone-Plan Subscription

STDS AT3

Project Report

Benedict Brunker: 25551995

Statistical Thinking for Data Science

TD School

University of Technology Sydney

STDS 36103

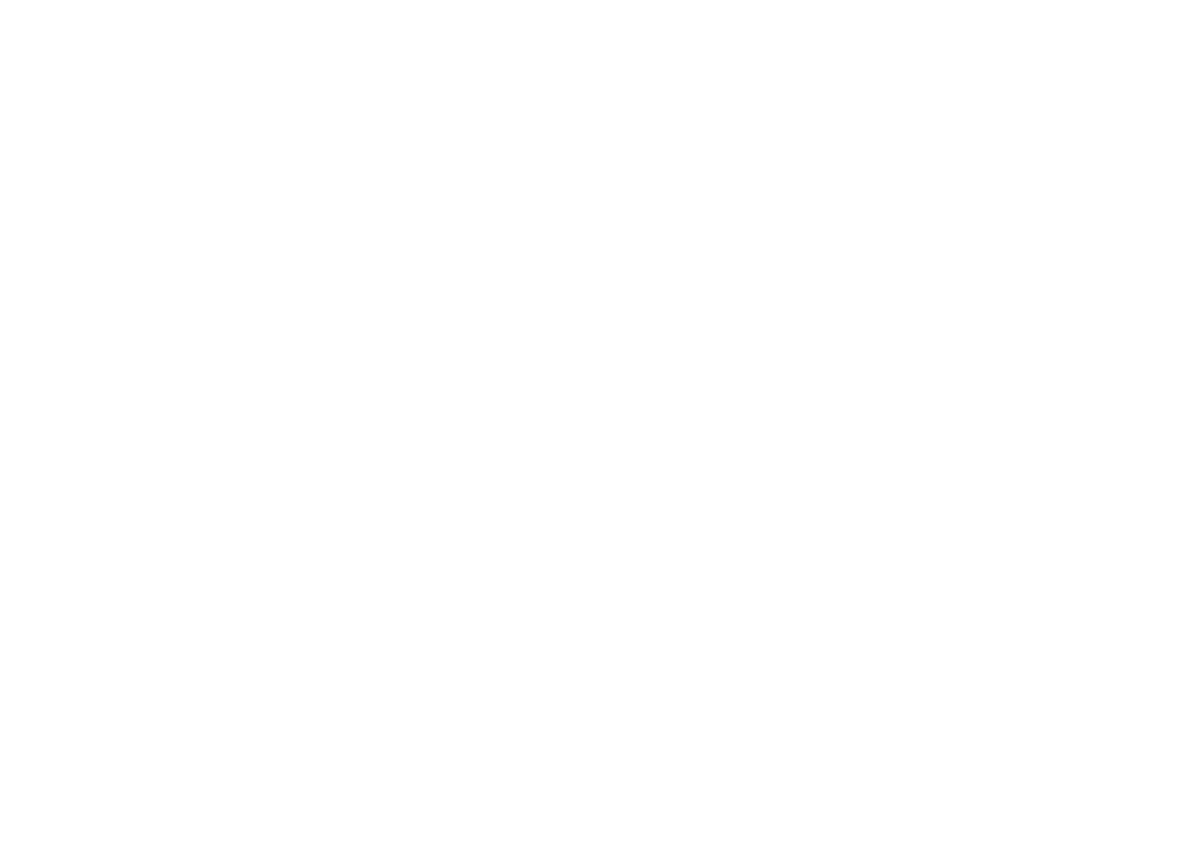
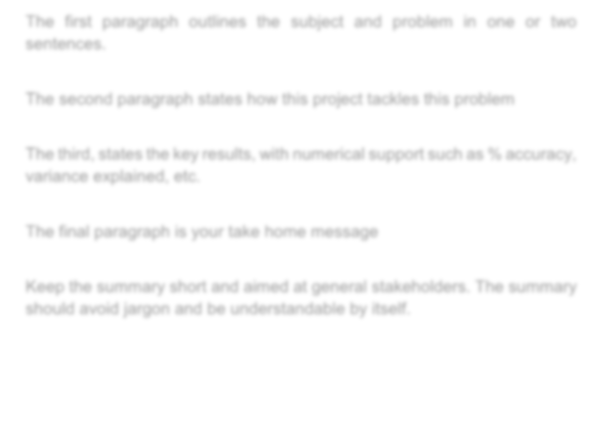
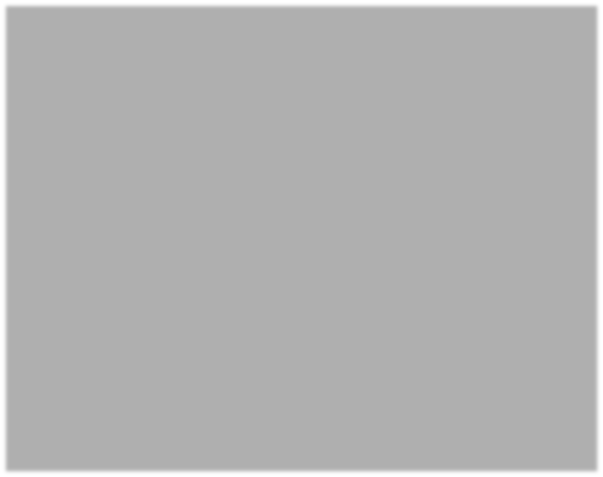
–

AT2B2 project report structure example



Executive

summary



STDS 36103 – AT2B2 project report structure example

# Table of Contents

**EXECUTIVE SUMMARY ............................................................................................................ 2**

[TABLE OF CONTENTS 3](#_Toc13713)

[INTRODUCTION 5](#_Toc13714)

[PROBLEM STATEMENT 5](#_Toc13715)

[RATIONALE 5](#_Toc13716)

[PROJECT AIMS AND OBJECTIVES 5](#_Toc13717)

[METHODOLOGY 8](#_Toc13718)

[METHODS OVERVIEW 8](#_Toc13719)

[METHODS DETAILS 8](#_Toc13720)

[RESULTS 11](#_Toc13721)

[KEY FINDINGS 11](#_Toc13722)

[IN-DEPTH RESULTS 11](#_Toc13723)

[Does increased smoking per captia cause a decrease in fertility rates? 11](#_Toc13724)

[CONCLUSION 14](#_Toc13725)

[TAKE HOME MESSAGE 14](#_Toc13726)

[DISCUSSION 14](#_Toc13727)

[PROJECT LIMITATIONS AND CAVEATS 14](#_Toc13728)

[STAKEHOLDER ANALYSIS AND PROJECT OUTCOMES 14](#_Toc13729)

[ACKNOWLEDGEMENTS 16](#_Toc13730)

[REFERENCES 17](#_Toc13731)

[APPENDICES 18](#_Toc13732)

STDS 36103

–

AT2B2 project report structure example



Introduction

STDS 36103 – AT3 – Predicting Subscriber Uptake

# Introduction

## Problem

The key to the long-term commercial viability of any telecommunications provider is its ability to sign up new subscribers. Mobile-phone plans are an inelastic service: an essential requirement for functioning in the modern world. But this also means that the market is heavily saturated: most consumers have a plan already. So, what entices consumers to change an existing plan? Under what conditions is this more or less likely? And what forms of promotion are most successful?

This project aims to address those questions through statistical learning. We build five types of classification models which will aid the business in understanding:

1. **Who** are the customers most likely to subscribe to a new plan, and what are their key characteristics?
2. **When** are the best conditions for successful campaigns?
3. **How** are new customers best reached?

STDS 36103

–

AT2B2 project report structure example



Methodology



STDS 36103 – Predicting Plan Uptake

# Methodology

## Modelling Approaches

Our experiments with statistical modeling can be grouped into two main categories:

1. **Parametric Models**: Logistic Regression (logit) models.
2. **Non-parametric Models**, including:
   1. Tree-based models, such as:

i. Simple Decision Trees (DTs)

ii. Random Forest Classifiers (RFCs)

iii. Gradient Boosting Machines (GBMs)

1. Distance-metric models, such as:

i.  K-Nearest Neighbors (KNN) Classifiers.

ii. Support Vector Classifiers (SVCs)

1. Neural Networks

## Meeting Parametric Assumptions

Logit models are “parametric” in that they assign coefficient weights to a given *p* number of parameters (ᵝ0 + ᵝ1 + … + ᵝp) (James et al., 2023). Logistic regressors make the following key assumptions about the data they model:

1. Independence of errors, essentially meaning that “all sample group outcomes are separate from each other (i.e., there are no duplicate responses)”.
2. “Linearity in the logit for any continuous variables”, meaning that higher or lower values for a continuous variable should increase or decrease the log-odds of the outcome.
3. An “absence of multicollinearity, or redundancy, among independent variables”.

(Stoltzfus, 2011, p.1101).

Exploratory Data Analysis (EDA) conducted in the first stage of this project showed the first assumption was satisfied (Brunker, 2024). The second assumption is satisfied by data transformation techniques described in Appendix B §3. As for the third assumption, EDA discovered two main areas of multicollinearity among variables:

1. **Temporal Variables** taking the form of daily, monthly, or quarterly macro-economic indicators. These naturally correlate both with one another, and with the variable ‘month’.
2. **Demographic variables** naturally correlate with one another, since they often function as components of *socio-economic classes* and *stages in the life-cycle.* For example, students are much more likely to be young (Quartile 1 for ‘age’), and much less likely to have mortgages (‘housing’ = ‘No’).

To deal with potential multicollinearity, we performed **Principal Components Analysis** (PCA) on these two variable types, yielding 3 *temporal components* and15 *demographic components.* This technique reduces covariance by condensing variables down into components which capture a maximum of explained variance in the data. This helps to satisfy parametric assumptions (1) and (2) above, since the resulting components take the form of a numerical continuous scale, which we could interpret as a kind of index and rename accordingly (cf. *Appendix B §3.6*).

**Logistic Regression**

We first designed three separate logit models fitted to variable types **a** and **b** above, along with a third, **contact variables**, which provide information as to *how* the customer had been previously contacted. Although the variable ‘duration’ (length of call with customer) is highly correlated with the target variable, we excluded this because the correlation is trivial; we cannot know how long a promotional call with a customer will last before we make the call, and our model is intended precisely to inform sales representatives as to which calls are best to make in the first place (Brunker, 2024). This exclusion makes models less predictive, but more useful in deployment.

There were three reasons for initially separating the variable types: 

1. To assess the predictive power of variable types on their own terms. This can be useful in deployment if, for example, we have information on macro-economic conditions but not customer profiles, or we want to know the likelihood of a customer subscribing across different macro-economic conditions.
2. To assess estimated coefficient weights and probabilities for variable types taken in isolation.
3. To determine the best selection of features for a comprehensive model that considers the three variable types together.

After building a comprehensive model through the above process, we tried both *Maximum Likelihood Estimation* (MLE) and *Bayesian Likelihood Estimation* (BLE) with five-fold cross validation to optimize coefficient weights.

Non-subscribers outnumber subscribers in the data by 4:1. For this reason, when splitting data into training, validation and testing sets, by a roughly 3:1:1 ratio, we stratified splits by the target class *subscribed.* We set the 𝜏 hyperparameter to 0.1, which instructs the model to predict 1 for any sample with 10% probability of subscription or higher. The rationale here is that we could justify 10 targeted outreach efforts for every new subscriber.

**Non-Parametric Models**

Because these models don’t make the same underlying assumptions about features as logit models, we could experiment with using both Principal Components (PCs) and the original variables as features. Generally we preferred PCs for more computationally expensive models and distance-metric models more likely to suffer from the “curse of dimensionality” (Hou et al., 2022) like Neural Networks, KNN and SVC, and preferred the original variables for tree-based models.

For these models we took three complementary approaches to tuning hyperparameters:

1. **Manual tuning** to grasp the relationship between hyperparameter values and results, and narrow the range of optimal hyperparameters.
2. **Grid-Search Cross Validation (GSCV)** to optimize hyperparameters within the discovered range.
3. **Bayesian Hyperpameter Optimization (BHO)** to compare against the GSCV optimized hyperparameters.

Where available, the ‘balanced’ setting for the *class\_weight* hyperparameter was useful in training models to be more sensitive to potential subscribers. Where unavailable, we used the Synthetic Minority Over-Sampling Technique (SMOTE) to generate “synthetic” duplicates of subscribing customers, allowing models to train on an even split of the target class.

STDS 36103

–

AT2B2 project report structure example



Results

STDS 36103 – AT2B2 project report structure example

# Results

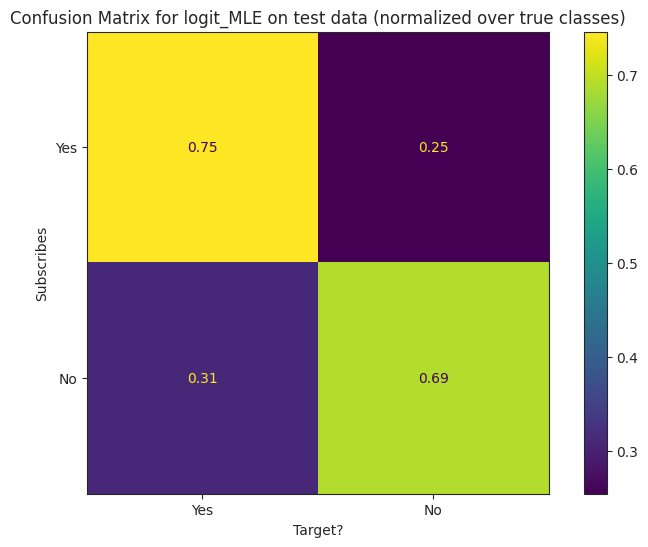
## Key findings

Put your one or two biggest, most important, or most surprising results here. Again, high level and high impact.

## In-depth results

Appendix *Tables A1-A3* shows the summary statistics for demographic, temporal principal components and contact variables estimated at less than a 5% chance of having no coefficient weight (p<|z|<0.05), while *Table A4* shows the same for the comprehensive model when uncertain features are excluded. *Figure 1* shows the Confusion Matrix for that model whose parameters are determined by five-fold cross validated Maximum Likelihood Estimation; the results when using Bayesian Likelihood Estimation were identical.

Figure 1



The logit model assigned the strongest positive coefficient weight to ‘cellular’, a Boolean variable indicating whether the customer was contacted by cell-phone (1) or landline (0). The strongest negative weight is estimated for ‘PC\_demo\_older\_studying’, a demographic component capturing older customers more likely to be in study. The smallest absolute weight was estimated for ‘pdays’ (number of days passed since customer was last contacted), which may be because our preprocessing for this variable failed to properly meet the linearity assumption. The most uncertain variable was ‘poutcome’ (previous outcome of the last campaign) with a p-value of 0.0952. ‘PC\_econ\_cool’, a component seeming to broadly capture ‘cool’ economic conditions of low employment, low inflation and low interest rates, was also somewhat uncertain at p=0.063, with a small negative coefficient weight of -0.0754. Overall, the contact variables and temporal components appeared more predictive than the demographic components.

**Non-parametric models**

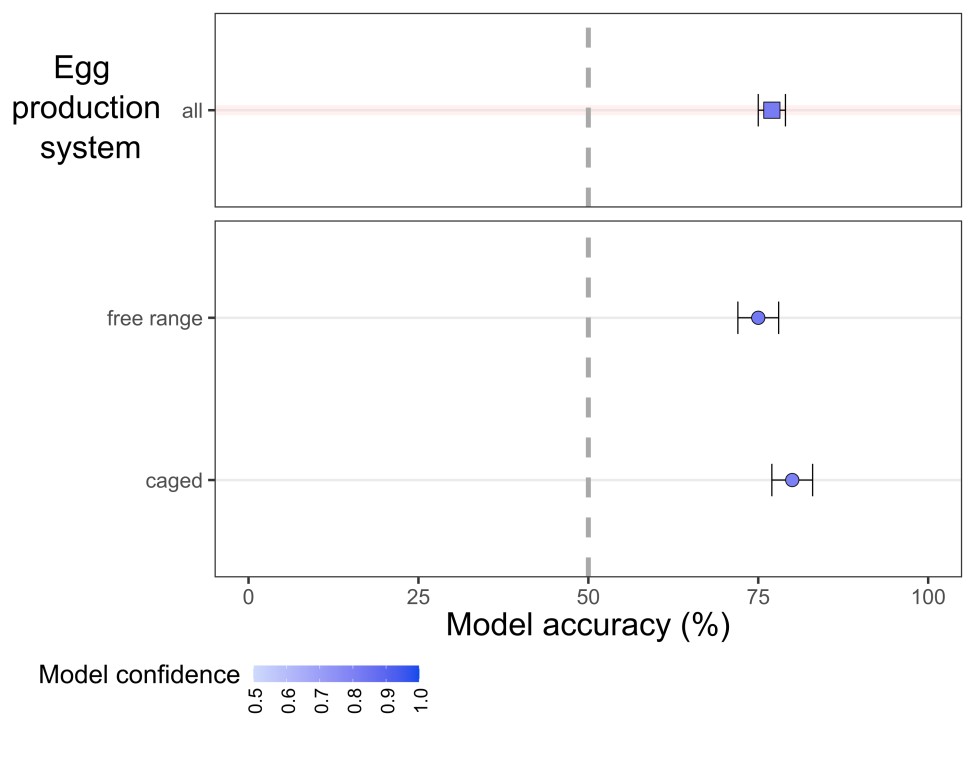
We tested the most promising model from each family; the comparative results are summarized in *Table 1* below. The neural network was the worst performer, showing no capacity for generalization by simply predicting 0 for all samples. This may be due to the tendency for neural networks to overfit, but may also owe to our own inexperience in dealing with their complexity. The other models yielded a broadly similar range of results but did not significantly exceed the performance of the logit model.

Use your research questions as subheadings for your results to help guide the structure of the results. Start with the biggest results and then go into details. Do not interpret your results too much here. You could say something like “Linear regression showed a significant negative slope between per captia smoking and fertility rate (slope = -2.074, p.value < 0.01) suggesting that smoking negatively impacts fertility rate”. However, you shouldn’t add additional speculation or synthesis here, do not, for example, say “This supports research by so-and-so 2015, and is likely due to smoking impacting hormonal… etc, etc”.

You can also highlight potential outliers here for later discussion. And should include a statement of how model assumptions were checked and if they were met. Diagnostic plots and coefficient tables should mostly be included in a separate appendix instead of the main text. You can refer to these as such “linear regression assumptions were checked via diagnostic plots and were determined to be met (Appendix A, figure 1)”.

In some cases you may have tried one method (Poission regression), but found an assumption wasn’t met (e.g. overdispersion), and switched to another method (negative binomial regression). You can and should mention this here to show that initial ideas were tested but needed changing.

Take care with your figure design. Figures are typically some of the first things that people check when looking at a document for the first time; poor figures leave a poor first impression that could be hard to recover from. Take the time and effort to make your figures pop and look pleasing. Send the figure to someone outside of the project and ask for their honest opinion, if they find it hard to read or that it looks ugly, chances are it is! Make sure text is easily visible and replace any shortened variable names with descriptive names with units: e.g. ‘fert’ to ‘Average fertility rate (births per capita)’. Try not to stick with ggplots default colours and theme, even though it shouldn’t make a difference, everyone knows what a default ggplot looks like, and it makes people assume you haven’t thought about the figure design. If you want more control over your figures, export them to .pdf and open in a vector graphics editor such as InkScape.



Overall

2019

2020

**Figure 1. Example results figure to show hypothetical results for model accuracy for predicting fertility rates from per captia smoking. Make your figure legends self-contained and descriptive so the reader can understand them. For example, overall accuracy was high suggesting that per capita smoking can be used to predict fertility rate. However, this figure doesn’t match with the research question. A better figure here would be a scatterplot showing the significant negative relationship between per captia smoking and fertility rate for all locations.**

STDS 36103

–

AT2B2 project report structure example



Conclusion



# Conclusion

## Take home message

What is the one thing--more than anything else in the world--that you want your reader to take away from your research? State that here with a follow up sentence about the implications or future work that this take home message inspires.

## Evaluation

We aimed to build classification models which powerfully predict customer subscription based on a set of *demographic, economic* and *contact* variables. The outcome was a success, but not an unqualified one. The overall accuracy of the best models was sound at over 90%, but we failed to reach the target we had set for ourselves in correctly identifying over 90% of potential new subscribers. Correct classification of potential subscribers was more important for this project than overall model accuracy, since the opportunity costs incurred by failing to subscribe new customers far outweighs the likely cost of targeted-advertisement per-customer. On the other hand, these results approximate those of the most analogous case-studies in the literature, such as that of Hou et al. (2021), which achieved better results overall, but at the cost of including the ‘duration’ variable, which we argue is only trivially predictive and therefore distorts the model.

## Reflections and recommendations for future work

Start the discussion by restating the aim of the project. If you achieved the aim, say so and provide evidence statements from your results. If you haven’t achieved your aim, provide a high level reason why you think this was the case before diving into more details further on.

The discussion should start specific and broaden back out. Think of it as the introduction in reverse; start with the project aims and objectives and then contextualise them in a broader context.

Try to thoughtfully connect the project outcomes with the wider themes underlying it. Smoking and fertility is one aspect of public health for example, what have you learned here that may have implications elsewhere.

Talk about any surprises and try to explain them. Things like outliers, or reversed hypotheses (what if smoking *increased* fertility rate in your study?). Whilst this is not a fully scientific article-style, try to look for and reference other relevant studies that support or maybe contradict your hypotheses or results.

## Project limitations and future recommendations

To improve on our own work, we would make the following suggestions for any future research in the field:

1. Our method of dealing with ‘unknown’ categorical responses by assigning them the value 0 as ternary variables was not clear, and the benefits of this approach were hard to assess. A better approach may have been to drop these observations (as did Hou et al., 2021).
2. Our attempt to deal with the class imbalance by using SMOTE seemed inadequate to improving sensitivity to the degree we had hoped. It is possible that SMOTE distorted models by training them on a large number of duplicates. It may be worth experimenting with Random Under-Sampling or some other method of addressing the class imbalance.
3. We may have failed to take full advantage of Neural Networks, mainly owing to a lack of expertise. In particular, it may be worth experimenting with Recurrent Neural Networks, which are better able to handle time-serialized features like ‘day’, ‘month’ and macro-economic indicators.
4. On a similar note, our way of handling temporal variables may have been sub-optimal. We excluded ‘day’ and ‘month’ from PCA since they seemed to confuse the results. However, it might have made more sense to try to co-ordinate daily indicators with the days themselves, monthly indicators with months, and quarterly indicators with financial quarters.
5. More data could always improve model performance, meaning both more observations and more variables. In particular, information about economic conditions and the kind of contact initiated with the customer seemed more predictive than demographic variables. A proper date-time variable would have helped in dealing with temporal variables described above.

Take time to be honest about the limitations of the study. This includes limitations in scope (amount of data, countries included, time periods included) and interpretation of outcomes (what your results do and so not tell you). When stating a limitation, it is important to appreciate and discuss it’s implications. For example, instead of just saying “One limitation was that all our data came from a single country, Australia.”, expand that, for example “Therefore we may expect our results to change if we included data from countries with lower GDP, because…, however, our results should hold across countries with similar socioeconomic conditions”.

## Key Insights

There are three key insights to draw from the project in answering the three main questions we posed in the *Introduction*. These insights are drawn from analysis of feature coefficients and importances.

1. *Who are our future customers?*

The models estimate that **students** are, *ceteris paribus,* mostlikely to subscribe to a new plan: being a student makes a customer roughly 90% to 196% more likely to subscribe than a non-student (within a 95% Confidence Interval). An effective future campaign might therefore consider marketing to the student population, perhaps by running promotions at Orientation Weeks or through Student Unions. Being a retiree is also strongly predictive of subscription, increasing probability of subscription between roughly 36.5% and 140% (again with 95% Confidence). Being ‘unemployed’ is also predictive of subscription, estimated at increasing the log-odds by 40.5%. On the other hand, holding a blue-collar job makes a customer ~16.75% less likely to subscribe to a new plan, likely because such jobs require daily use of a mobile phone.

Divorcees appear to be least likely to subscribe to a new plan, and singletons most.

The demographic profile of customers **most** likely to subscribe are those in **secure self-employment.** These are customers who are more likely to report as ‘self-employed’, less likely to have ‘housing’ or personal loans, are middle-aged and of average education (see *Figure 3*).

On the other hand, the demographic profile of customers **least** likely to subscribe are

Here you can state what the key project outcomes are and the specific relevant stakeholders. Project outcomes can be an artefact (e.g. a regression model that can be used for prediction) or knowledge (e.g. a newly discovered relationship between your response and explanatory variables). Project outcomes are not

coefficient tables, data visualisations, or p.values. Think of an outcome as something someone else could use in some way.

Once your outcomes are defined, you should then state the key stakeholders that you envisage could benefit from it. If it is a predictive model, it could be a tool governments could use to make predictions for future planning. If it is knowledge, it could be something used by policy makers to inform the public or help draft legislation.

Outcomes and stakeholders do not need to be a huge list, identify the most important outcomes, and for those outcomes the main stakeholders.

# Acknowledgements

Typically here is where you can mention other people or organisations outside of the main project team that supported throughout the project. Probably not applicable here given that you are all working in teams with no engagement outside of that, but it is good to be aware of.

# References

Hou, S., Cai, Z., Wu, J., Du, H., & Xie, P. (2022). Applying Machine Learning to the Development of Prediction Models for Bank Deposit Subscription. International Journal of Business Analytics (IJBAN), 9(1), 1-14. <http://doi.org/10.4018/IJBAN.288514>.

# Appendices

**Logit Coefficient Weights by Maximum Likelihood Estimation**

You can include additional tables and figures as appendices to support the report. This should not be an infinite stream of every figure or analysis you can think of, but should include relevant additional information. Examples would be regression assumption plots (Q-Q, residuals) to support assumptions. It may include regression tables or statistical tests showing if normality has been met, or overdispersion tests for a Poission regression.

Everything here should serve some supporting purpose for the project. Make sure to check and justify as a team what goes into the appendices. Also, do not include additional text here outside of figure and table legends. All additional text will be ignored for marking the main report.