

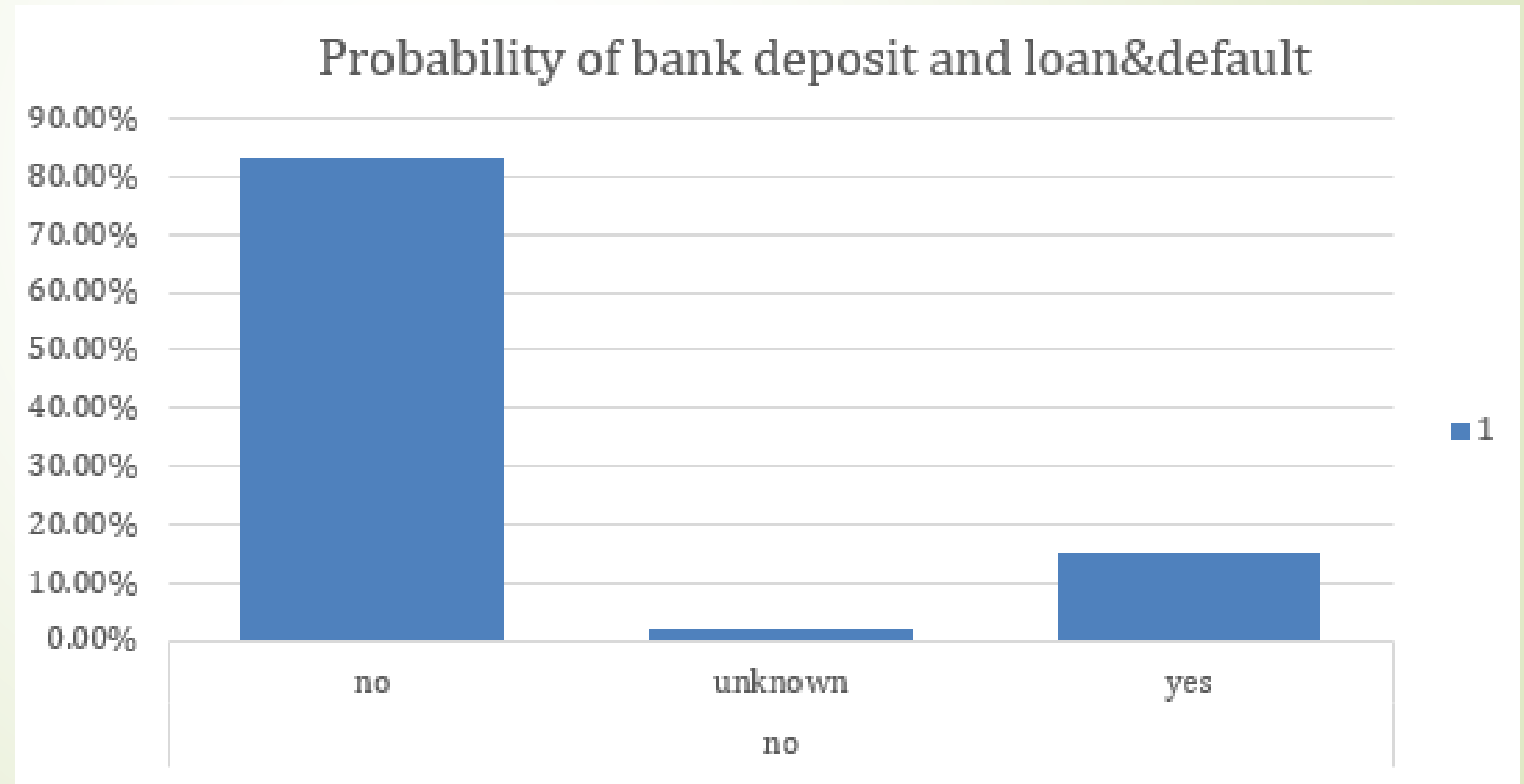


ETC3250 – Project Presentation

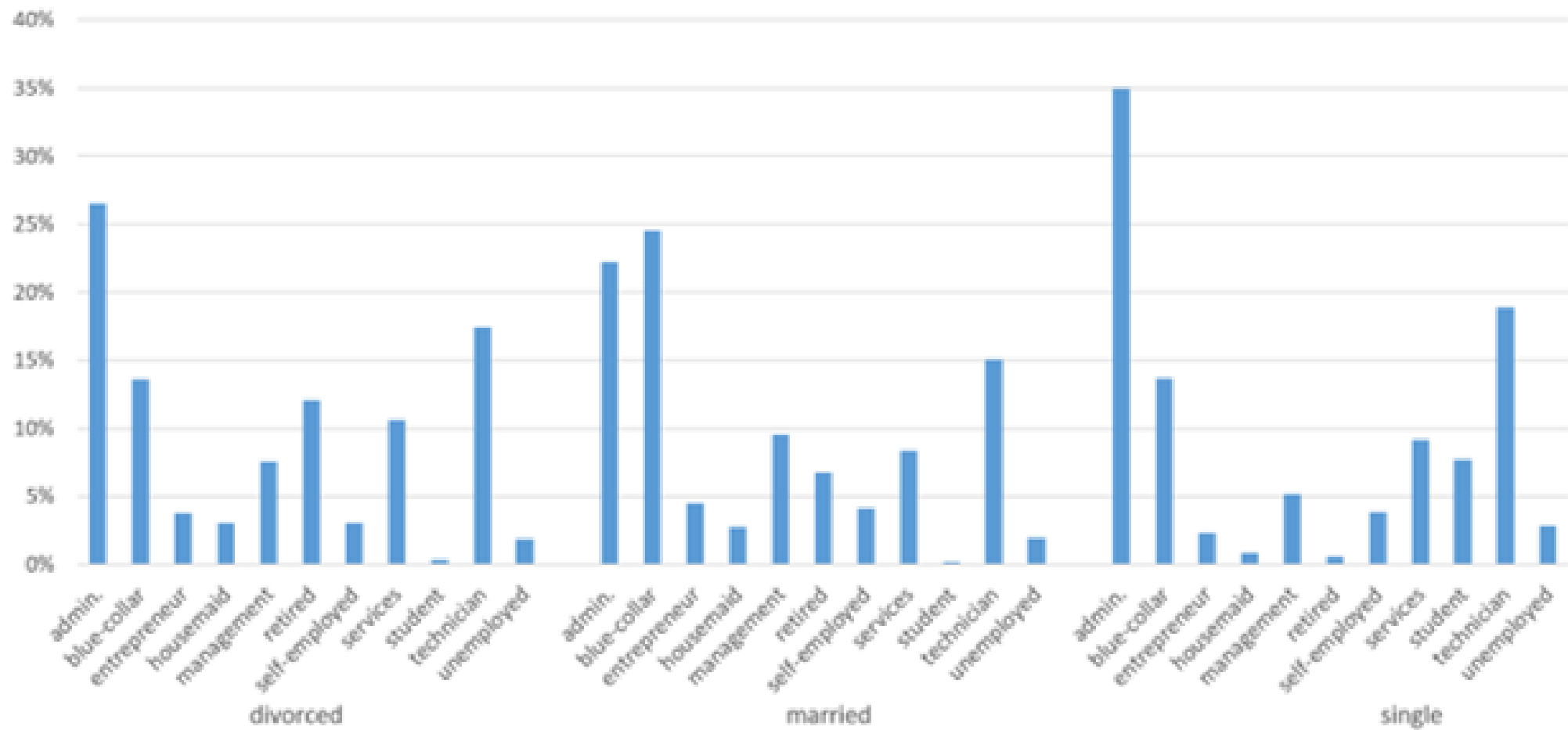
William Chan (25961039), Yung Chyi Siah (27717518), Connor Lickliter (27794628), Mihir Bhatt (25175319), Alec Kajewski (26895765)

Introduction

- To produce the most accurate prediction model for how likely it is a client will subscribe to a bank term deposit.
- 7 predictors and 7 variables related to previous campaign



Probability of bank deposit and marital&job status



Methodology

- Initially the data set was examined to show the range/concentration of values and to identify potential outliers
- Various methods were used including logistic regression, probit models and decision trees
- Initial logistic regression:
 - $\hat{Y} = a\hat{g}e_i + j\hat{o}b_i + m\hat{a}r\hat{i}t\hat{a}l_i + e\hat{d}_i + h\hat{o}u\hat{s}i\hat{n}g_i + l\hat{o}a\hat{n}_i$



admin.	blue-collar	entrepreneur	housemaid	management	retired
7496	7054	1175	858	2233	961
self-employed	services	student	technician	unemployed	unknown
1099	2999	305	5280	718	258

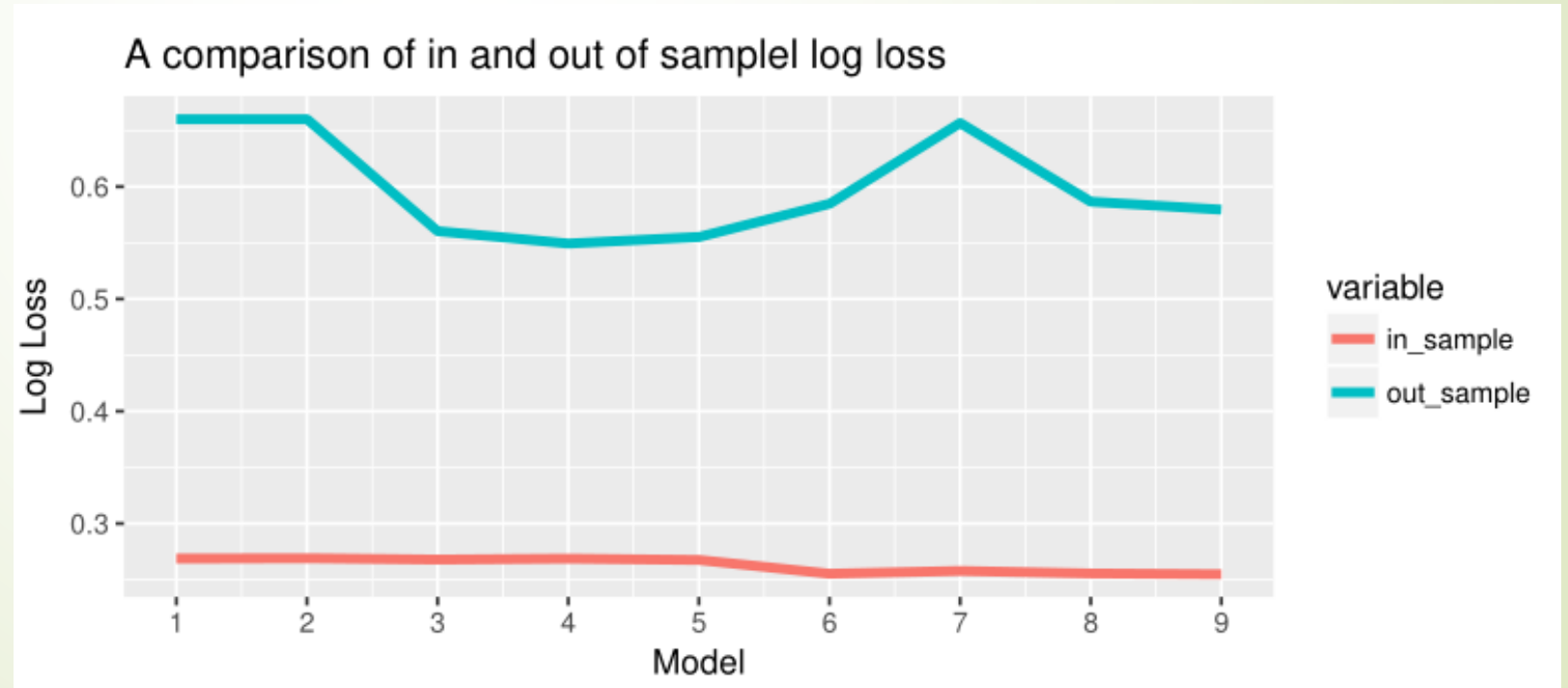
Methodology

- The predictions from the model were then calculated
- To compare model predictions, the in sample log loss was calculated using the following function in R

```
log_loss = function(prediction) {  
  log_loss = -mean(my_data$y*log(prediction) + (1 - my_data$y)*log(1-prediction))  
  return(log_loss)  
}
```

Results and Discussion

- Evaluation Method
 - In Sample Log loss
 - Overfitting



Results and Discussion

- Data set manipulation and reduction
- fit = **glm**(y ~ retiredstudent + single + poutcome_binary, data = my_data, family = binomial) (In sample log loss = 0.26835 vs out of sample = 0.55529)

```
Coefficients:
              Estimate Std. Error z value Pr(>|z|)
(Intercept)   -2.61321    0.02654  -98.473  < 2e-16 ***
retiredstudent  0.76599    0.08278   9.253  < 2e-16 ***
single         0.25854    0.04731   5.465 4.63e-08 ***
poutcome_binary 1.98677    0.20530   9.678  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

- Investigation into best model
- fit = **glm**(y ~ marital + default + loan + pdays + poutcome, data=my_data, family=binomial) (In sample log loss = 0.26879 vs out of sample = 0.54941)

Results and Discussion

- Tree functions
- Random Forests

Alternative Methods

