**SLIDE 1:**

* Marketing has become the heart of a business corporation.
* Crucial that the obtaining of consumer information through promotion and telemarketing is given a great deal of importance.
* In the case of bank institutions, marketing often consists of drawing clients to specific investment opportunities or offering long-term deposits contracts.
* Purpose of the analysis: Discover the factors that positively and negatively affect the probability of any given client subscribing to the ‘Bank Term Deposit’ offered by the bank institution.
* From a business perspective: Understand what factors impact the decision of a customer, i.e., to subscribe to the bank institution’s term deposit or not.
* Understanding what factors impact this decision = bank can fine-tune its marketing effort to target the demographics that correlate with higher probabilities of subscription.
* Thus, saving time and money for a business, which is a recipe for the success of any business.

**SLIDE 2: DATASET**

* Dataset of Portuguese Banking Institution involving direct marketing campaigns from 2008-2013.
* From UCI Machine learning Rep., 40,000 rows, 20 variables.
* Marketing campaigns performed by phone call. More than 1 method of contact often used, in order to determine if a client was to be subscribed to a “Bank Term Deposit” (‘yes’ or ‘no’).

**SLIDE 3: EXPLORATORY ANALYSIS**

* RIGHT GRAPH: # of ‘yes’ increases as duration increases, which makes sense because the longer a client is willing to hear about an offer, their interest is heightened.
* LEFT GRAPH: it is seen that retirees, unemployed, students, and management position clients are all more likely to sign up for long term deposits, as the proportion of yes to know is greatest for these demographics.

**SLIDE 4: MAIN DATA ANALYSIS**

TOP GRAPH: ‘housing’

* Blue is the same for each bar in the graph, showing the irrelevance of the housing variable, and for unknown, housing is a null value, not informative.
* Thus, having a house loan is seen to be insignificant.

BOTTOM GRAPH: ‘day\_of\_the\_week’

* Same idea for each bar, so the day is irrelevant for yes vs. no, and it is seen that the # of contacts are close for every given day, which shows that days do not influence the decision. Also seen to be insignificant.

**SLIDE 5: INITIAL MODEL**

* Standard errors of the coefficients are very large for this model. This is likely due to overestimation. Definitely not a great fit.
* However, difference in deviance shows the absence of important

predictors that are not present in the dataset.

* For predictors we do have, it does answer some questions.
* Stepwise was performed to produce an AIC from the full model being 17,694 to the final model being 17,685. This is a marginal improvement.
* Important variables determined to define the response are clearly lacking. But given the fit we have, we have tangible results such as the coefficients.
* CHART: All p-values of the resulting model’s predictors are significant, and thus, this is the best model.

**SLIDE 6: SIGNIFICANT COEFFICICENTS**

* Holding all other variables fixed, a 1-minute increase in duration results in a 0.5% increase in the odds of a subscription.
* Holding all other variables fixed, a unit increase in consumer confidence index results in a 2.1% increase in the odds of a subscription.
* Holding all other variables fixed, if the customer is a student, the odds of subscribing increase by 29%. This is in line with our initial hypothesis.
* Holding all other variables fixed, for every unit increase in the number of times a customer is contacted, there is a 16.9% increase in the odds.

**= These are important takeaways from our analysis**

**SLIDE 7: MODEL PERFORMANCE**

* Assessing the performance of the model can see that AUC value is high which shows good performance.
* Success in partitioning of the response variable.
* Surprising that optimal threshold is unexpected at 0.085, meaning that anything above 0.085 would be a ‘yes’.
* Suspected to be BECAUSE class distribution is very unbalanced (Count for ‘yes’ vs. ‘no’ results), hence it is very low.

**SLIDE 8: CLASSIFICATION TABLE**

* As seen by the 304 **false-predicted ‘no’s’** and the 6957 **false-predicted ‘yes’**, it is seen that there are far more false’s than true’s. ACCEPTABLE.
* Overall, the tradeoff is **acceptable** because you would rather advertise to people that are not interested in the product than not advertise to people that are, in fact, interested.

**SLIDE 9: GOODNESS OF FIT**

* J = 10 partitions, and the result did not change for a different J (5-20).
* Hosmer-Lemeshow statistic following a Chi-Squared distribution 571.12. This x-squared value is high given the number of degrees of freedom in the model.
* The null hypothesis is that there is no lack of fit. P-value of 2.2x10-16 proves a rejection of the null, proving there is a lack of fit with our model, as suspected.

**SLIDE 10: CONCLUSION**

* Graph 1: Observed proportion is linear estimator results divided into quantiles. The number of ‘yes’ responses are divided by the total number of observations per quantile.
* Predicted probability does not line up properly with the observed proportions.
* Predictions are lower than the observed, which can be traced back to lack of a balanced class distribution.

Graph 2: Residuals vs. Linear Predictor

* Residuals are not even in their variance with respect to the linear predictor.
* All of this justifies a model that does not correctly fit the data.

**SLIDE 11: DISCUSSION**

* Overall, the huge difference in deviance from full and null show that there is an error in specification, indicating a clear lack of fit.
* Class distribution makes sense to be a big reason for this lack of fit. There are much more ‘no’s than ‘yes’s, which definitely indicates an imbalance in class distribution. It is tough to gain insight from data when the responses are so skewed towards one specific response. Perhaps if there were more yes’s, it would be easier to interpret the true influence of predictors to the extent of designing a well-fit model.
* Perhaps there is omitted variable bias in the model, and thus there are missing variables that are crucial in predicting the response for our model.
* Or, perhaps, applying transformations to some predictors would improve the model fit.
* Perhaps the relationship we’ve tried to capture is not linear and thus, it may be feasible to include higher order terms.
* Additionally, another option would be considering that maybe there is a better model choice. Such as an ADDITIVE MODEL because of its added flexibility in interpreting non-linear relationships like this one may be.