Draft2

Motivation

- **Bank marketing dynamics (cite the paper)**

Bank telemarketing campaigns are a major channel through which financial institutions promote term deposit products. Thsee marketing campaigns come with a cost and are often inefficient due to low client response rates. Prior work by Moro et al. (2014) discover that client characteristics and campaign related factors strongly influence subscription outcomes. Motivating a data-driven approach to improve targeting and decision-making in bank marketing. In their approach a Portuguese retail bank was studied using campaign data collected between 2008 and 2013.

- **Introduce dataset**
  - Description of variables (under factors(
  - Description of target variable (EDA of several factors with target; plots)
  - note imbalance
  - Include wrangling of dataset (as.factor() and addition of variable for pdays)

Our dataset was a publicily available daaset, thus had less variables for privacy concerns of the bank customers. The Bank Marketing dataset was collected by a Portuguese banking institution during a series of direct telemarketing campaigns. These campaigns were done to encourage clients to subscribe to a long-term deposit product. Telemarketing is an expensive and time-consuming process; therefore, the bank would want to know in advance which customers are most likely to subscribe before any more investment takes place. The dataset consists of sixteen variables describing demographic, financial and campaign related variables that enlist individual clients contacted during direct marketing efforts. A list of the 16 variable found below...

- age: Age in years
- job: Occupation (Categorical)
- marital: Marital status (Categorical)
- education: Highest level of education (Categorical)
- default: has credit in default? (binary)
- balance: average yearly balance
- housing: has housing loan? (binary)
- loan: has personal loan? (binary)
- Contract: contact communication type (Categorical)
- day: last contact day (of the month)
- month: last contact month of year (month; 1=January)
- duration: last contact duration, in seconds
- campaign: number of contacts performed during this campaign and for this client
- pdays: number of days that passed by after the client was last contacted from a previous campaign (-1 means client was not previously contacted)
- previous: number of contacts performed before this campaign and for this client
- poutcome: outcome of the previous marketing campaign (Categorical)
- y: Indicator of whether the client subscribed a term deposit (Binary target)

The response variable indicates whether a client subscribed to a term deposit which defines a binary classification task.

Exploratory analysis shows that the numeric predictors (balance, duration, pdays, previously contacted) are heavily skewed. Also, analysis reveals substantial class imbalance, in the response variable, which motivates the use of appropriate modeling and evaluation for our classification problem. Visualizations of subscription rates by job education marital status contact type and previous outcome reveal variations in client responsiveness across different factors.

[STREAMLINE Plots and EDA summary HERE]

Data preprocessing involved converting categorical variables to factors and adding derived variables such as pdays to account for clients not previously contacted. Converting the categorical variables (initilay character type) to factors creates an indicator for each class in the variable (i.e. if job has 20 factors, 20 indicator variables will be considered in the model fit). As for the previously contaced variable (days since previously contacted), customers who haven't been prreviously contacted will recieve a value of -1. This can cause improper modeling of customers, since the variable now takes on the value of negative numbers. Instead we set -1 to zeros and created a new variable to indicate customers that have not been previously coontacted before. These steps ensure that the dataset is suitable for statistical learning models and enable accurate assessment of predictive performance.

- **Problem Statement**
  - introduction of ML (as in ISLR)
    - Chapter 2: Estimating f(x)
      - How some models are more flexible, or more accurate. How some are less interpretable
      - Assessing model accuracy
    - Chapter 4: We are dealing with classification problem

This study is motivated by the availability of bank marketing datasets that enable comparison of machine learning methods for predicting term deposit subscription behavior.

The specification of the problem begins with the recognition that we are addressing a binary classification task where the goal is to predict whether a client will subscribe to a term deposit based on demographic financial and campaign features. According to the framework presented in Introduction to Statistical Learning (ISLR), the fundamental goal is to estimate the unknown function, f(x) that relates the predictors, X to the response, Y. Some models are highly flexible and can capture complex non-linear relationships between predictors and the response. In most Cases these models achieve higher predictive accuracy. However, these flexible models often suffer from lower interpretability, making it difficult to understand the influence of individual features. Simpler models such as logistic regression are less flexible but provide coefficients that can be directly interpreted in terms of odds ratios, offering insight into feature importance.

Chapter 5 of Introduction to Statistical Learning emphasizes the importance of obtaining an unbiased estimate of a model's test error to assess its predictive performance on unseen data. Test error quantifies how well a model generalizes beyond the data used for training, which is critical for selecting among competing models. The chapter presents several methods for estimating test error, including the simple training/test split, K-fold cross validation, and the leave-one-out cross validation (LOOCV) approach. Each method balances bias and variance differently, with cross validation generally providing a lower-variance estimate compared to a single split. These techniques are essential for model selection, tuning hyperparameters, and comparing the expected performance of alternative models. For the Bank Marketing dataset, estimating test error accurately is particularly important due to the class imbalance and the presence of both categorical and numeric predictors, which can influence model generalization.

In the following chpater of Introduction to Statistical Learning focuses on model selection and regularization techniques to improve predictive performance and reduce overfitting. It introduces methods such as subset selection, shrinkage (Ridge and LASSO regression), and dimension reduction to assess and enhance model

accuracy while controlling variance. These approaches allow practitioners to identify parsimonious models that generalize well to new data, providing complementary strategies to traditional test error estimation. We apply these concepts to the models we fit on the bank data, to reduce overfitting and ultimately improve the generalizability of our models to new clients.

`- Models we will implement (without diving into details)`

We start with a logistic regression (described in methodology section), then build up adjustments and regularizations to the logestic regression to make our baseline. We then move to another interpretable model: KNN. This model is considered transperent in a unique, due to it's neighboring factor. Finally we fit a Boosting model. This state-of-the-art model is not considered interpretable. Although, there are methods used to extract the contributing variables in such models. We included this model is used to gage how much accucracy we are sacrificing for interpretability.

`- Bank marketing dynamics (cite the paper) differentiate our methodology`

The authors employed time ordered data splitting rolling window evaluation. For feature enrichment, they implimented a semi-automatic approach including external intuitive knowledge from domain experts (bank manager) to optimize features. They compared Logistic regression with Decission trees, SVMs, and Neural networks; all complex 'black-box' models. While Moro et al. prioritized maximizing classification performance, our approach also examines model simplicity, automatic feature selection, and diagnostic checks for appropriate classifications (from confusion matrices).

Methodology

For each of our models we implement a k-fold CV approach, with k=5 training/test split to maintain the distribution of the response variable across the sets. Our models do not apply time-series modeling, since our data does not include sequintial variables. Although variables such as campaign month or day are available within the dataset, which can be considered to hold temporal patterns, we do not model this data as a time series. This keeps the focus on cross-sectional relationships between client characteristics and subscription behavior.

1. **Logistic regression**

Logistic regression is used to model the probability of a binary outcome as a function of predictor variables. The model estimates the probability that a client subscribes to a term deposit, $P(Y = 1 \mid X)$. This model fits with assumptions of independent observations and normality.

Formula:

$$f(x) = P(Y = 1 \mid \mathbf{x}) = \frac{1}{1 + e^{-(\beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k)}}$$

Coefficients are estimated using the maximum likelihood method, which finds the parameter values that maximize the likelihood of observing the training data (ISLR).

Feature selection is performed using backward stepwise selection to identify a parsimonious subset of predictors that minimizes estimated test error as measured by AIC. Variables such as duration and previous marketing outcomes remain in the model due to their strong predictive power, whereas less informative variables are excluded to reduce variance. Categorical predictors are converted to factors, and a new binary variable previously_contacted is created to handle clients with no prior campaign contact. This preprocessing ensures that the data is suitable for statistical modeling and facilitates accurate interpretation of model coefficients. We resulted 12 features:

```
selected_formula
```

```
## y ~ job + marital + education + balance + housing + loan + contact +
##     day + month + duration + campaign + poutcome
```

Cross validation is used to select optimal tuning parameters for these regularized models and to estimate the expected test error. Confusion matrices, sensitivity, specificity, and ROC/AUC metrics are computed to assess predictive performance, especially given the class imbalance where non-subscribers are the majority. This model performed will with mean 5-fold CV test error metrics:

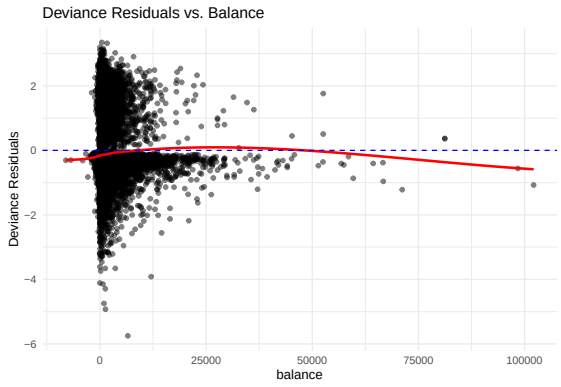
```
print(conf_matrix)
```

```
## Confusion Matrix and Statistics
##
##           Reference
## Prediction   no  yes
##        no  7786  710
##        yes  198  347
##
##                Accuracy : 0.8996
##                  95% CI : (0.8932, 0.9057)
##     No Information Rate : 0.8831
##     P-Value [Acc > NIR] : 3.496e-07
##
##                   Kappa : 0.3842
##
##  Mcnemar's Test P-Value : < 2.2e-16
##
##             Sensitivity : 0.32829
##             Specificity : 0.97520
##          Pos Pred Value : 0.63670
##          Neg Pred Value : 0.91643
##              Prevalence : 0.11691
##          Detection Rate : 0.03838
##    Detection Prevalence : 0.06028
##       Balanced Accuracy : 0.65174
##
##        'Positive' Class : yes
##
```

Next, we check for potential non-linearity in continuous predictors from our predictors remaining after back-ward stepwise selection. We plotted deviance residuals against numeric features. Based on residual diagnostics, we incorporated a polynomial term (using the poly function: *poly(balance, 4)*) for balance to account for non-linear effects. The Balance squared and power 4 term produced significant coefficients.

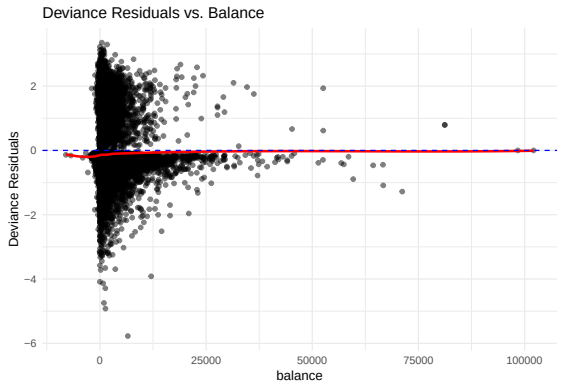
```
p1
```

```
## `geom_smooth()` using formula = 'y ~ x'
```




```
p2
```

```
## `geom_smooth()` using formula = 'y ~ x'
```




```
tran.coef
```

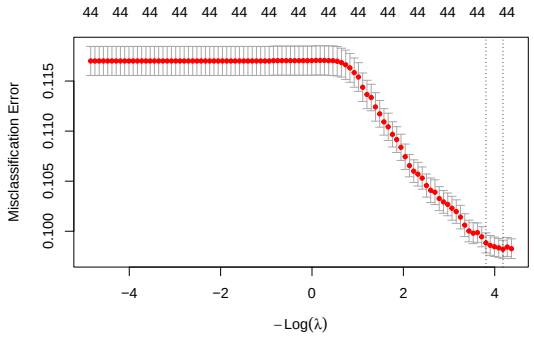
```
##                          Estimate   Std. Error     z value      Pr(>|z|)
## (Intercept)         -2.620324324 1.462577e-01 -17.9158087  8.876960e-72
## jobblue-collar      -0.341741366 8.082951e-02  -4.2279283  2.358529e-05
## jobentrepreneur     -0.434778497 1.431727e-01  -3.0367419  2.391501e-03
## jobhousemaid        -0.471871035 1.494571e-01  -3.1572345  1.592732e-03
## jobmanagement       -0.218611280 8.164726e-02  -2.6775090  7.417186e-03
## jobretired           0.244849687 9.781371e-02   2.5032245  1.230675e-02
## jobself-employed    -0.264167753 1.232030e-01  -2.1441663  3.201956e-02
## jobservices         -0.249710190 9.367276e-02  -2.6657717  7.681185e-03
## jobstudent           0.406418753 1.193512e-01   3.4052348  6.610719e-04
## jobtechnician       -0.230498448 7.673979e-02  -3.0036367  2.667737e-03
## jobunemployed       -0.281480151 1.267009e-01  -2.2216106  2.630963e-02
## jobunknown          -0.529337133 2.711010e-01  -1.9525461  5.087341e-02
## maritalmarried      -0.136229076 6.650159e-02  -2.0485084  4.051021e-02
## maritalsingle        0.152626849 7.123702e-02   2.1425216  3.215153e-02
## educationsecondary   0.198757856 7.250831e-02   2.7411736  6.122016e-03
## educationtertiary    0.376357446 8.383368e-02   4.4893345  7.144604e-06
## educationunknown     0.243747043 1.168074e-01   2.0867432  3.691136e-02
## housingyes          -0.658643127 4.866831e-02 -13.5333055  9.944923e-42
## loanyes             -0.423872041 6.753970e-02  -6.2758944  3.476302e-10
```

```
## contacttelephone    -0.097396340 8.257965e-02  -1.1794230  2.382298e-01
## contactunknown      -1.542551296 8.151426e-02 -18.9236978  7.276559e-80
## day                  0.010571185 2.802164e-03   3.7725081  1.616147e-04
## monthaug            -0.688040538 8.792122e-02  -7.8256485  5.050465e-15
## monthdec             0.592969598 2.009252e-01   2.9511962  3.165457e-03
## monthfeb            -0.155244505 1.003194e-01  -1.5475019  1.217423e-01
## monthjan            -1.283009396 1.370470e-01  -9.3618221  7.837323e-21
## monthjul            -0.843818467 8.669431e-02  -9.7332622  2.175035e-22
## monthjun             0.384506815 1.050647e-01   3.6597158  2.524952e-04
## monthmar             1.618134654 1.350791e-01  11.9791680  4.568858e-33
## monthmay            -0.421069070 8.094596e-02  -5.2018543  1.973099e-07
## monthnov            -0.821506981 9.252600e-02  -8.8786605  6.767032e-19
## monthoct             0.825399254 1.188899e-01   6.9425501  3.850842e-12
## monthsep             0.847194731 1.326409e-01   6.3871315  1.690263e-10
## duration             0.004185982 7.244492e-05  57.7815759  0.000000e+00
## campaign            -0.080669584 1.112260e-02  -7.2527663  4.083443e-13
## previous             0.008839947 6.503592e-03   1.3592406  1.740703e-01
## poutcomeother        0.246298284 9.968511e-02   2.4707631  1.348251e-02
## poutcomesuccess      2.337670803 8.898360e-02  26.2708055 4.135594e-152
## poutcomeunknown     -0.063946934 6.711537e-02  -0.9527912  3.406959e-01
## poly(balance, 4)1    5.598179724 4.351371e+00   1.2865323  1.982574e-01
## poly(balance, 4)2  -23.024570764 7.714952e+00  -2.9844089  2.841268e-03
## poly(balance, 4)3   -0.913732363 7.930483e+00  -0.1152177  9.082725e-01
## poly(balance, 4)4  -19.804601172 6.260653e+00  -3.1633445  1.559676e-03
```

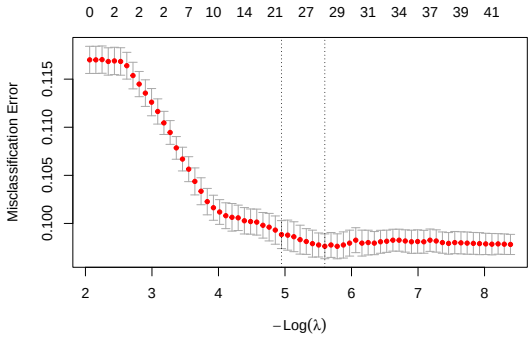
Although this polynomial transformation didn't provide desired metrics in expense of the nonlinear effect. From the deviance residual plots we can observe there weren't significant deviance from a normal residual line.

Regularization techniques such as Ridge (l2 penalty) and LASSO (l1 penalty) are employed to reduce variance and perform feature selection, particularly in the presence of correlated predictors. LASSO (Least Absolute Shrinkage and Selection Operator) regression uses an ℓ_1 penalty ($\alpha = 1$). A key property of the LASSO is that it forces the coefficients of some variables to be exactly zero, thereby performing automatic feature selection (Chapter 6). The LASSO plot similarly uses cross-validation to select the optimal $\lambda_{\min}$, 0.0037. Unlike Ridge, the LASSO is expected to yield a more parsimonious model by excluding irrelevant predictors entirely.

```
plot(cv.ridge)
```



```
plot(cv.lasso)
```

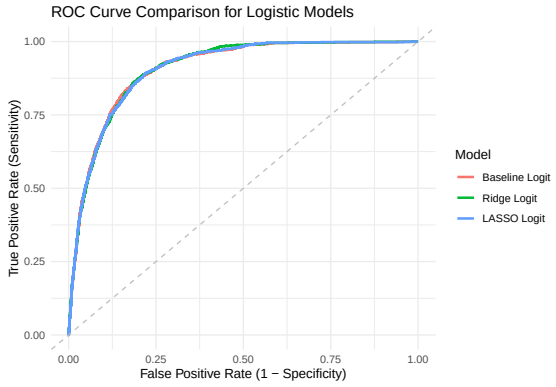


Here is a comparison of the models fitted thus far:

```
metrics
```

```
##            Model  Accuracy Sensitivity Specificity
## 1 Baseline Logit 0.8996792   0.3301798   0.9750752
## 2    Ridge Logit 0.8973565   0.2686850   0.9805862
## 3    LASSO Logit 0.8981307   0.3008515   0.9772044
```

Ridge and LASSO models yield similar results, indicating that additional shrinkage does not meaningfully improve classification of subscribers. We assume this is due to our modeling being majority indicator variables. This binary aspect of the indicator variable causes it's coefficient to be less affected by regularization techniques.



All three models exhibit a high overall Accuracy (around 0.89), which is expected given the significant class imbalance (Prevalence ≈ 0.11). However, the critical metric, Sensitivity (the True Positive Rate for the minority class, 'yes'), is very low (around 0.17). This result, consistent across all three models, indicates that they struggle to correctly identify subscribers, leading to many False Negatives (FN). Conversely, the high Specificity (True Negative Rate ≈ 0.98) means the models are excellent at identifying non-subscribers. The regularization methods (Ridge and LASSO) failed to provide any significant improvement in these metrics, suggesting that the primary limitation is one of high bias (due to insufficient flexibility) rather than high variance.

2. **KNN**

- Describe fitting algorithm (from ISLR)
- sample code and plots form model_fit file
- short paragraph after each action ("we did this because..." or "we didn't do this because...")

3. **Boasted trees**
    - Describe fitting algorithm (from ISLR)
    - sample code and plots form model_fit file
    - short paragraph after each action ("we did this because..." or "we didn't do this because...")

Results and Discussion

- Declare we are looking for specificity and interpretable (and cite back to specification of problem from introduction)
- Stats
    - Table of metrics for all models
    - plot AUC curve
- Business impact
    - Demonstrate interpretable models benefits
- Future work

Our methodology differentiates from prior work by Moro et al. (2014) in focusing on a comparative evaluation of interpretable models with minimal feature engineering, emphasizing a balance between predictive power and model transparency.

Model performance was primarily assessed using the area under the ROC curve which provides a threshold independent measure of classification quality. In contrast our work focuses on the original bank marketing dataset and emphasizes comparative evaluation of standard statistical learning models with reduced feature complexity and greater interpretability.

References

1. dataset
2. Paper on website
3. ISLR v2