Exam PA June 18, 2020 Project Solution

Instructions to Candidates: Please remember to avoid using your own name within this document or when naming your file. There is no limit on page count.

Also be sure all the documents you are working on have June 18 attached.

As indicated in the instructions, work on each task should be presented in the designated section for that task.

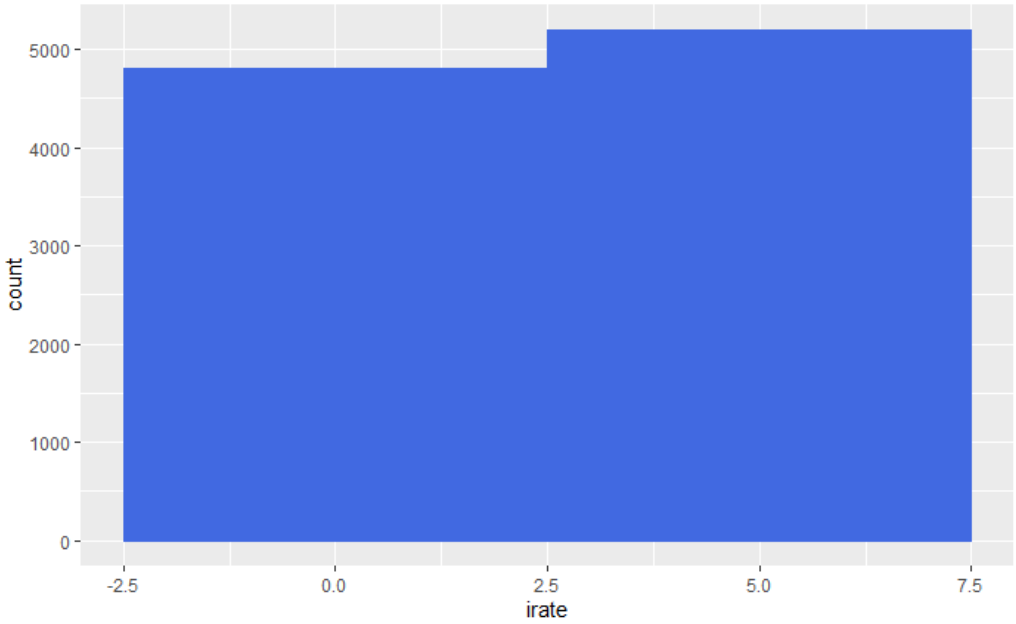
> *This model solution is provided so that candidates may better prepare for future sittings of Exam PA. It includes both a sample solution, in plain text, and commentary from those grading the exam, in italics. In many cases there is a range of fully satisfactory approaches. This solution presents one such approach, with commentary on some alternatives, but there are valid alternatives not discussed here.*

Task 1 – Explore the data (8 points)

Candidates were expected to analyze and comment on a variety of charts, but some candidates did not comment on all of them, losing credit. Many candidates did not add a comment about how it would impact future modeling but only observed how it would look statistically. That was the biggest challenge for candidates on this task.

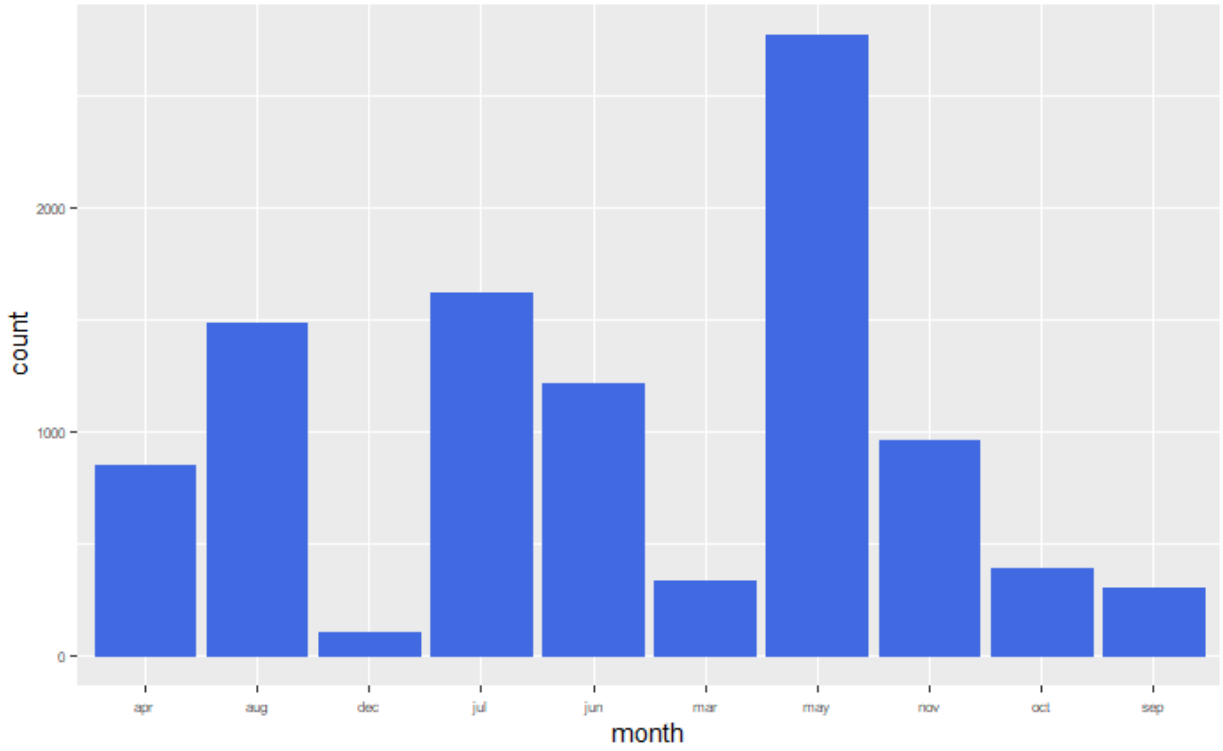
Irate

Clearly the assistant made poor choices in building this chart, but candidates were not expected to alter them, just comment on them as best they could.

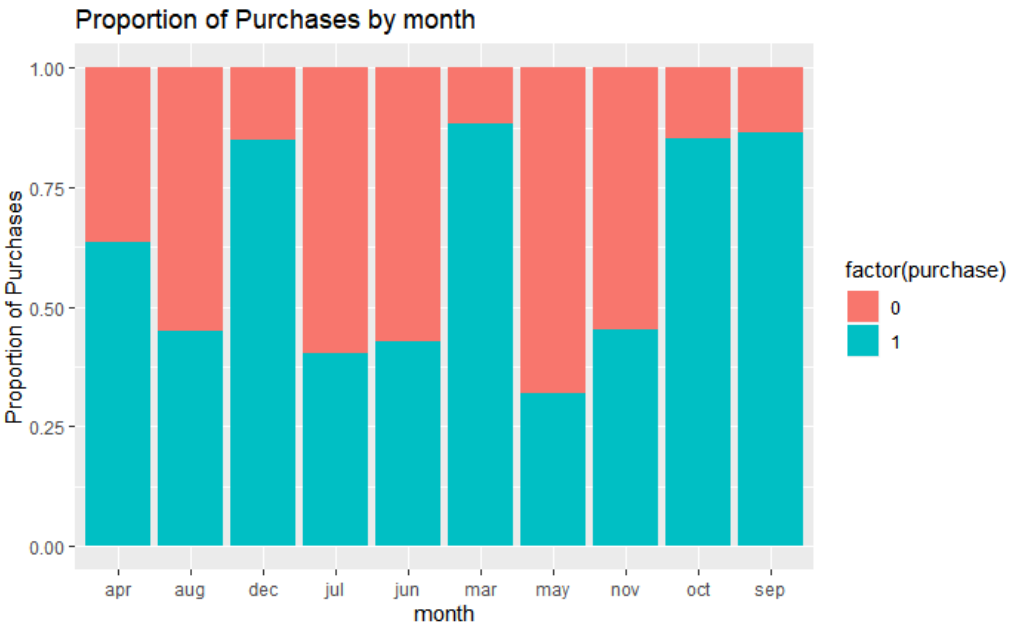


The histogram for irate doesn't reveal much information as shown, just that a similar number of records occur for irate above and below 2.5. It is not clear whether negative interest rates values exist—this should be checked later for reasonability.

Month

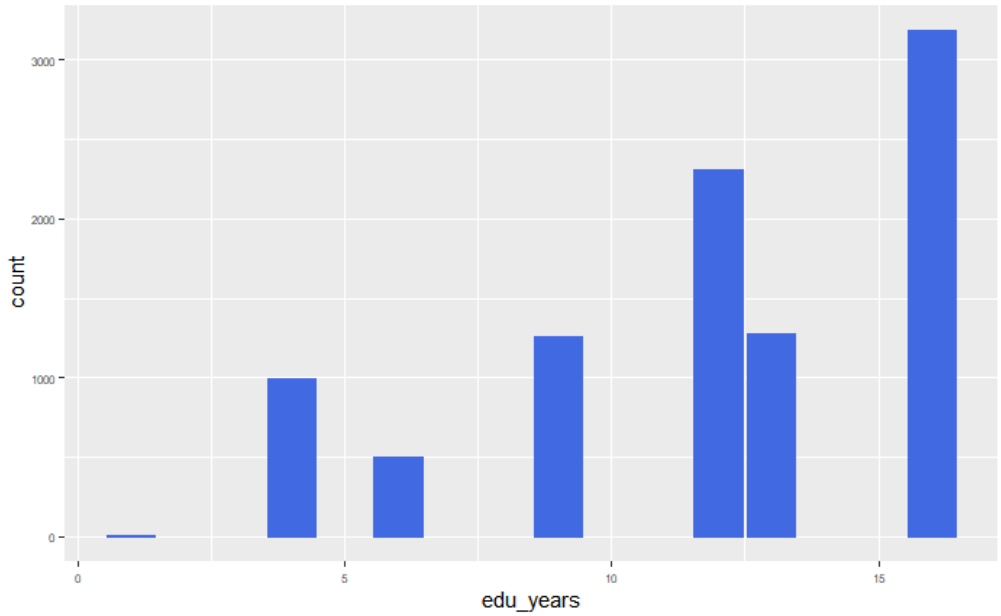


The most frequent month is May, followed by the summer months of July, August, and June. There is no data for January or February, but model predictions may be needed for these months, with a decision to be made regarding whether these should be like neighboring months or the baseline month. Also, the categories are ordered alphabetically when ordering by month would be more sensible for understanding results.



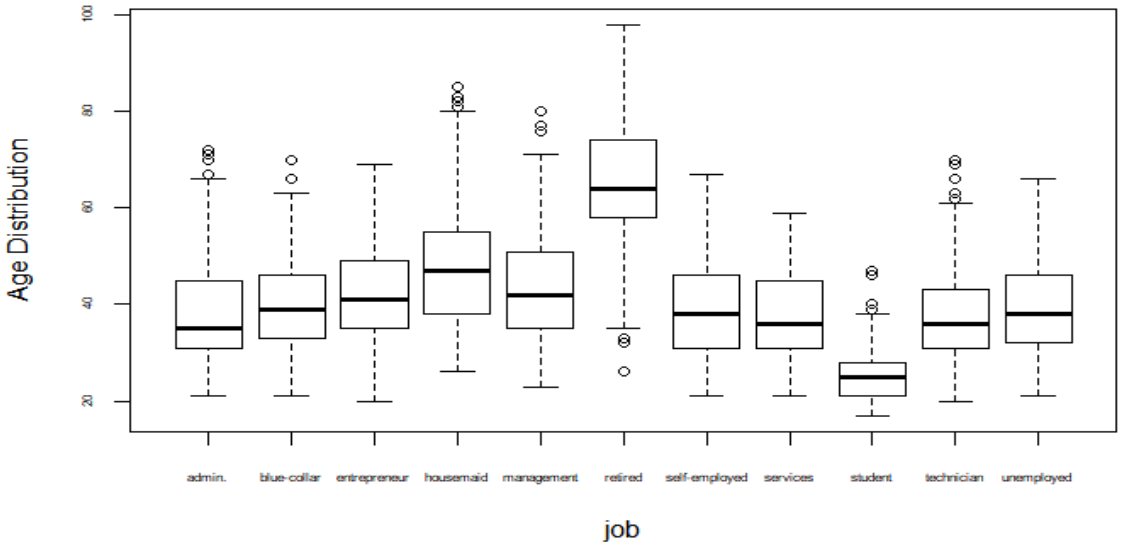
The most frequent months also have the lowest purchase rates. While this may be attributable to these being summer months, it may also be the case that higher volumes of calls are less productive in themselves and these higher volumes just happened to be in the summer months. The relationship between density of calls and purchase rate should be considered when modeling.

Edu_years



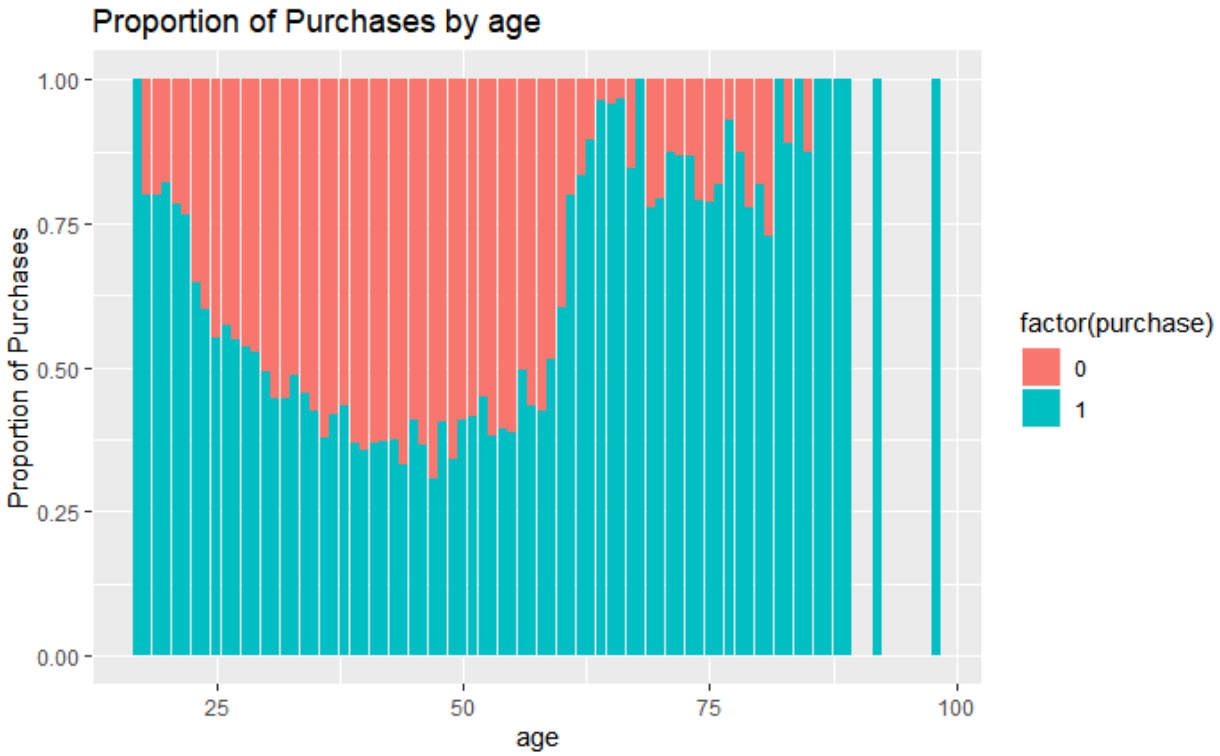
Edu_years takes integer values from 1 to 16, with noticeable gaps between values. It may be difficult to decide how to assign missing values given this distribution. Also, there are very few observations having a value of 1, an outlier that may need to be removed given its distance from other observations.

Age Distribution by Job



In looking at age by job, two categories, retired and student, pop out immediately. It makes sense that retirees tend to be at older age while students are more likely to be at younger age. As age and job have some codependence, we need to be careful in dealing with these two variables together in the same model. Particularly for greedy decision trees but also possible for GLM, strong results for one variable may hide the influence of the other variable.

Age



The proportion of purchase has a downward trend roughly between age 17 and age 50 and then starts increasing after that, particularly a significant jump around age 60 and thereafter. GLM models will have trouble fitting this down-up curve in purchase by age with only a single age variable, and additional variables based on age, for example age squared, will be needed to capture the observed trends for GLM. Decision trees will not need additional variables to capture such shapes.

Task 2 – Consider the education variable (3 points)

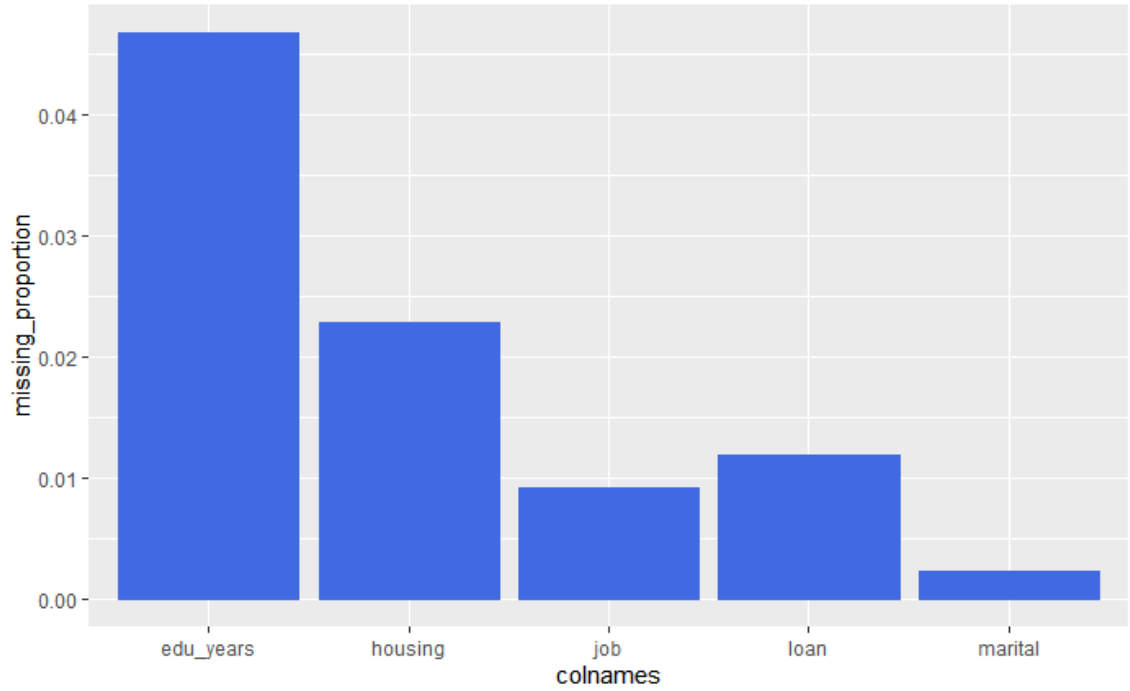
Few candidates considered the difference in dimensionality from using a numeric variable. Often, candidates did not distinguish between GLM and decision trees. Some candidates, when discussing decision trees, compared numerical and categorical and said one was better than another without recognizing that, in this case, the decision tree could ultimately reproduce the original categories.

A strategy for dimension reduction is necessary to avoid the curse of dimensionality, which can lead to overfitting. The original categorical variable on education requires six variables, in addition to a baseline category, in the model. On the other hand, creating a new numeric variable, edu_years, only requires one variable. While some information is lost in going from six variables to one, the risk of overfitting can be greatly reduced depending on the model used.

GLM models can better discern linear trends over several ordinal categorical variables, as seen in edu_years, when they are converted to numerical variables. For decision trees, however, the conversion to numerical does not reduce the dimensionality as much because the tree can still split between any adjacent pair of variables and, with enough splits, reproduce the categorical variables.

Task 3 – Handle missing values (5 points)

The edu_years variable presents a challenge as each method for removing missing data has material flaws. Overall, alternative approaches for variables could often be justified.




```
[1] "Purchase Proportions by variable, for missing and non missing values"
[1] "  Variable      PP_for_NAs  PP_for_non_NAs"
[1] "   housing            0.47            0.46"
[1] "       job            0.41            0.46"
[1] "      loan            0.46            0.46"
[1] "   marital            0.52            0.46"
[1] " edu_years            0.54            0.46"
```

Edu_years: impute using mean. Almost 5% are missing, and the purchase proportion is significantly higher for missing values than non-missing values. Removing these may cause us to lose valuable insights. Being a numeric variable, converting to an "unknown" value does not work, so imputing the missing value using the mean is left, though we do not have confidence that what causes these to be missing is spread evenly among education.

Housing: convert to "unknown". As over 2% are missing, converting to unknown despite not much difference in purchase proportions between missing and non-missing values.

Job: convert to "unknown". Less than 1% are missing but the purchase proportion is noticeably less than that of non-missing values and adding one more category to this high-dimension variable will not hurt the modeling much.

Loan: remove rows. Just over 1% are missing and the purchase proportion of these is nearly identical to that of non-missing values, so it is not worth adding a third category.

Marital: remove rows. So few rows are missing that the higher proportion is not likely to be predictive if converted to "unknown", so better to remove instead.

Task 4 – Investigate correlations (3 points)

Which correlations are concerning can be a matter of judgment, though the 94% correlation between irate and employment clearly needed to be discussed. Most candidates did not relate concerns on the correlations to specific modeling techniques, which was needed to earn full credit. Some candidates mentioned clustering as an alternative technique, but the type of clustering was important, as some clustering approaches do not easily accommodate new observations for prediction.

	age	edu_years	CPI	CCI	irate	employment
age	1.00000000	-0.23508994	-0.01709708	0.14426794	-0.04064880	-0.07196693
edu_years	-0.23508994	1.00000000	-0.08549192	0.03923499	-0.07728819	-0.07759744
CPI	-0.01709708	-0.08549192	1.00000000	-0.13962175	0.58691796	0.37381941
CCI	0.14426794	0.03923499	-0.13962175	1.00000000	0.06115118	-0.07637779
irate	-0.04064880	-0.07728819	0.58691796	0.06115118	1.00000000	0.94220249
employment	-0.07196693	-0.07759744	0.37381941	-0.07637779	0.94220249	1.00000000

The most notable correlations are:

- irate and employment (0.94)
- CPI and irate (0.59)

These correlations and others are not that concerning for decision trees. For example, irate and employment are heavily correlated, and so no or little information can be gained from splitting on employment after having split on irate and the second variable will be excluded. The only concern is that the variable chosen may flip-flop depending on training data and the modeler may not be aware that the other is almost as predictive.

These correlations are concerning for GLM models, which do not handle highly collinear variables well. Very large and mostly offsetting coefficients may result, making interpretation of the coefficients difficult. In particular, it is dangerous to interpret the coefficient as representing the impact on the target variable with other variables held constant, given that the correlated variable is likely to also change. The accuracy of the estimated coefficients is also questionable and different results can occur if a new sample is taken.

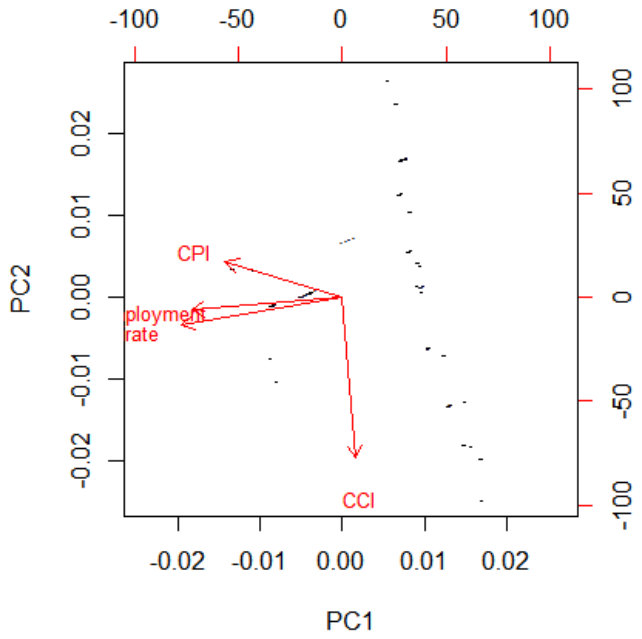
One method other than PCA for handling the correlated variables is to use one of the variables and delete the redundant ones.

Task 5 – Conduct a principal components analysis (7 points)

Candidates generally printed the bi-plot but often did not indicate how to read the plot when explaining the loadings. Few candidates interpreted the plot very well. The explanations for why it is appropriate not to consider age and edu_years varied widely in quality.

A good way of handling correlated variables is to perform principle components analysis (PCA) to obtain

orthogonal variables in which different principle components are uncorrelated, but still containing most of the information. Here, I did a PCA on the following variables: CPI, CCI, irate, and employment. I set the scale parameter to "TRUE" so variables can be scaled to have unit variance. Without it, certain variables could dominate the associations between the variables due to larger magnitudes of variance.



In the PCA bi-plot, the relative loadings as seen in the red scales and arrows are of the most interest for comparing the first two principle components, PC1 and PC2. Employment and irate have nearly identical positions, showing that PC1 and PC2 do not distinguish much between them. In PC1, similar movements in these two variables and CPI are grouped together with little emphasis on CCI, while PC2 highlights movements in CCI, combined with some opposing movement in CPI. The variation from PC2 is visible in the black PC scores, with a wide (tall, really) variation of PC2 scores for PC1 scores between 0.01 and 0.02.

Age and edu_years will not be helpful additions to the PCA. Their correlations to the other variables are weak. As a result, their inclusion will make it more difficult to interpret the principal components while likely decreasing the proportion of variance explained by the initial components. Including them in the PCA is unlikely to reduce the number of dimensions used in the GLM without additional information loss compared to not including them.

Task 6 – Create a generalized linear model (5 points)

While most candidates had little trouble with the output of the model, some candidates struggled with the explanation on the differing performance of the age variable. The ROC curves and AUC were given as a convenience to help recognize the difference in model performance, but other ways could have been used to compare the models.

To begin, an age only GLM with a logit link was built on the training data set.

```
Call:
glm(formula = purchase ~ age, family = binomial(link = "logit"),
    data = data_train)
```

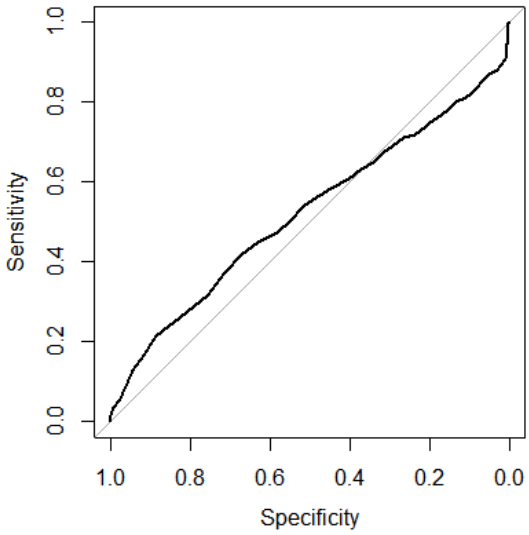
```
Deviance Residuals:
   Min      1Q  Median      3Q     Max
-1.237  -1.105  -1.072   1.255   1.321

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept) -0.451800   0.085438  -5.288 1.24e-07 ***
age          0.007126   0.002024   3.520 0.000431 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 9520.8  on 6900  degrees of freedom
Residual deviance: 9508.4  on 6899  degrees of freedom
AIC: 9512.4
```

Based on the output shown above, the age variable has a low p-value, showing that, in isolation, it is a statistically significant predictor.



In addition, the area under the above ROC curve (AUC) for the test data is 0.5210. Clearly, this age only single factor model is not doing very well. Its performance is little different from the 0.5 expected for an intercept-only model. The curving around the diagonal shows that while its predictions at higher ages (corresponding to higher probabilities of purchase in the model) are slightly helpful, its predictions at lower ages, as seen at the top right, miss that purchase rates are higher again at the very lowest ages.

Next, we ran a full GLM model with a logit link on the following variables.

- age
- job
- marital
- edu_years
- housing

- loan
- phone
- month
- weekday
- PC1

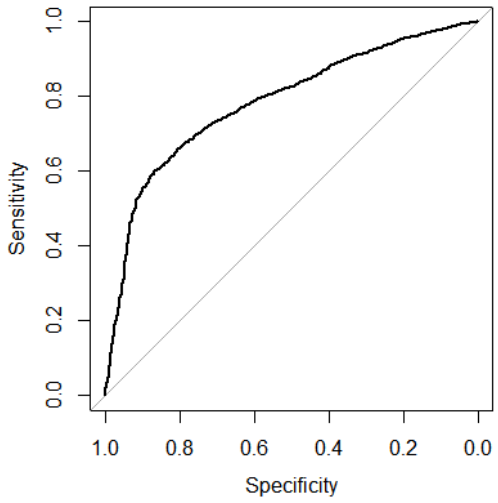
```
Call:
glm(formula = purchase ~ age + job + marital + edu_years + housing +
    loan + phone + month + weekday + PC1, family = binomial(link = "logit"),
    data = data_train)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.5048  -0.8604  -0.5524   0.8524   2.2025

Coefficients:
                  Estimate Std. Error z value Pr(>|z|)
(Intercept)      -0.792982   0.256324  -3.094 0.001977 **
age               0.002162   0.003305   0.654 0.513077
jobblue-collar   -0.052183   0.101739  -0.513 0.608015
jobentrepreneur  -0.101667   0.161324  -0.630 0.528559
jobhousemaid      0.176139   0.189823   0.928 0.353454
jobmanagement    -0.047579   0.117922  -0.403 0.686599
jobretired        0.422056   0.158505   2.663 0.007751 **
jobself-employed  0.091358   0.164583   0.555 0.578838
jobservices      -0.067070   0.109165  -0.614 0.538954
jobstudent        0.405282   0.168806   2.401 0.016356 *
jobtechnician     0.018179   0.088193   0.206 0.836691
jobunemployed     0.181155   0.184333   0.983 0.325724
jobunknown       -0.489341   0.362874  -1.349 0.177494
maritalmarried    0.098413   0.092947   1.059 0.289687
maritalsingle     0.113271   0.106451   1.064 0.287295
edu_years         0.024965   0.009714   2.570 0.010171 *
housingyes       -0.043629   0.056781  -0.768 0.442262
housingunknown    0.046082   0.266910   0.173 0.862927
loanyes          -0.169497   0.077422  -2.189 0.028579 *
phonelandline     0.124089   0.088910   1.396 0.162811
monthaug          0.174566   0.122880   1.421 0.155426
monthdec          0.427699   0.326408   1.310 0.190087
monthjul          0.639901   0.126692   5.051 4.40e-07 ***
monthjun          0.443279   0.126786   3.496 0.000472 ***
monthmar          1.011835   0.210983   4.796 1.62e-06 ***
monthmay         -0.635487   0.103780  -6.123 9.16e-10 ***
monthnov         -0.041248   0.126637  -0.326 0.744636
monthoct          0.885805   0.200839   4.411 1.03e-05 ***
monthsep          0.853477   0.233275   3.659 0.000254 ***
weekdaymon       -0.066205   0.090445  -0.732 0.464169
weekdaythu        0.072307   0.088349   0.818 0.413115
weekdaytue        0.102365   0.091600   1.118 0.263773
weekdaywed        0.111862   0.090875   1.231 0.218344
PC1               0.681394   0.028945  23.541  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 9520.8  on 6900  degrees of freedom
Residual deviance: 7691.4  on 6867  degrees of freedom
AIC: 7759.4
```



This model is performing much better with a test data AUC of 0.7830, showing that the model makes some effective predictions due to the additions of many other predictors. Given the much better AUC, it does not feel like age is among the more robust predictors. Age now has a high p-value and is no longer in itself a good predictor. Age is not independent from other variables, particularly job, and the trends ascribed to age in the age-only model are much better described by other variables, for instance "job: retired", once they are included in the model.

Task 7 – Select features using stepwise selection (8 points)

Few candidates explained well why adding the square of age as a feature may improve the model despite the clear quadratic-looking curve given in task 1.

Some candidates skipped the discussion on best subset selection—those who answered it mentioned efficiency, appropriately. Fewer noted how stepwise processes may not find the best model.

Many candidates did not explicitly list the variables chosen as asked for in the question.

Adding the square of age as a feature to improve the model is a reasonable suggestion. As seen in the exploratory data analysis in Task 1, both younger and older ages had higher purchase rates while middle ages had lower purchase rates. If age, but not its square, is included, the signal seen for either younger or older ages will be lost completely—the model will not be capable of producing a non-linear response to age. Adding the quadratic term gives the model more flexibility to fit the shape observed in Task 1, though with a slightly higher risk of overfitting the data.

For the purpose of feature and model selection, one can use best subset selection, fitting separate GLMs for each possible combination of features and then select the best combination. However, when there are many predictors in a model, the number of possible combinations can be very large, making best subset selection impractical and computationally inefficient.

An alternative is stepwise process, constructing a model by adding (forward selection) or removing (backward selection) one predictor at a time. It is a simpler and faster process compared to best subset selection, but it may not find the optimal combination of features.

Using backward selection and AIC, rather than forward selection and/or BIC, is the pair of choices that will need to more variables being retained in the model. Since our goal in this project is to identify the key variables that relate to the target variable, this approach may not do enough selection of variables to help identify the most important variables as too many variables will remain. This is not a reasonable approach given the business problem.

The summary report of the resulting model can be found below.

```
Call:
glm(formula = purchase ~ age + I(age^2) + edu_years + loan +
    phone + month + PC1, family = binomial(link = "logit"), data =
data_train)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.7203  -0.8611  -0.5578   0.8552   2.1221

Coefficients:
                Estimate Std. Error z value Pr(>|z|)
(Intercept)    0.9732981  0.3511868   2.771 0.005581 **
age           -0.0790796  0.0153816  -5.141 2.73e-07 ***
I(age^2)       0.0009261  0.0001723   5.375 7.68e-08 ***
edu_years      0.0268703  0.0077480   3.468 0.000524 ***
loanyes       -0.1802954  0.0771393  -2.337 0.019425 *
phonelandline  0.1289511  0.0885259   1.457 0.145213
monthaug       0.2025489  0.1216448   1.665 0.095896 .
monthdec       0.4312462  0.3263844   1.321 0.186407
monthjul       0.6178053  0.1261329   4.898 9.68e-07 ***
monthjun       0.4329255  0.1257552   3.443 0.000576 ***
monthmar       0.9701195  0.2106397   4.606 4.11e-06 ***
monthmay      -0.6326691  0.1029112  -6.148 7.86e-10 ***
monthnov      -0.0279457  0.1260064  -0.222 0.824485
monthoct       0.8860034  0.2002607   4.424 9.68e-06 ***
monthsep       0.8555358  0.2330934   3.670 0.000242 ***
PC1            0.6705547  0.0288847  23.215  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 9520.8  on 6900  degrees of freedom
Residual deviance: 7690.8  on 6885  degrees of freedom
AIC: 7722.8
```

The final model from stepwise selection has picked the following variables:

- Age
- Age^2
- Edu_years
- Loan
- Phone
- Month
- PC1

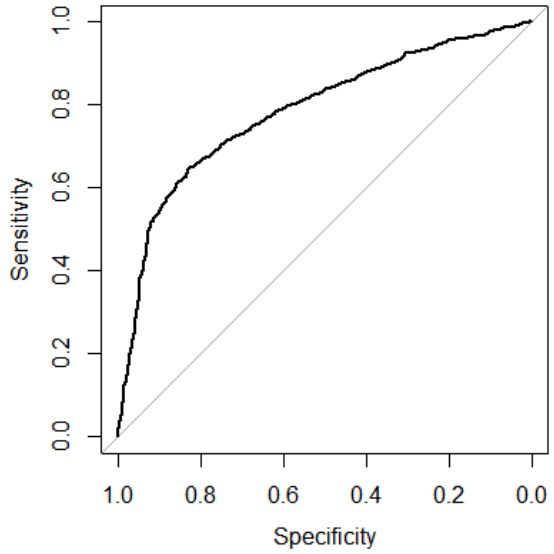
Task 8 – Evaluate the model (12 points)

On AUC, most candidates described AUC near 1 well but many struggled with explaining 0.5 (which is not just half right and half wrong) and near 0. Typically, an AUC of less than 0.5 would not happen and would indicate a problem in the model optimization or metric calculation.

Many candidates did not remember to compare variables selected to previous tasks as stated in the problem statement.

Many candidates did well to provide observations to marketing beyond numerical interpretation, making them more discussable and actionable.

The final model from Task 7 has AUC's of 0.7786 and 0.7842 on training and testing, respectively, indicating the model fit is good overall. The ROC curve for the test data is shown below.



A perfect model that predicts the correct class for new data each time will have a ROC plot showing the curve approaching the top left corner with an AUC near 1.0. When a model has an AUC of 0.5, like when the ROC curve runs along the diagonal shown, its performance is no better than randomly selecting the class for new data such that the proportions of each class matches that of the data. Any model having an AUC less than 0.5 means it is providing predictions that are worse than random selection, with a near 0 AUC indicating that the model makes the wrong classification almost every time.

In the data exploration, age and month were expected to have an impact on the proportion of purchase, and this has been confirmed by the final model from Task 7 due to the inclusion of the age squared term, as the second GLM in Task 6 did not find age to be a reliable predictor when age squared was absent. Compared to the second GLM in Task 6, only 6 of the 10 variables are retained after feature selection was applied, and age squared was added, but the dropped variables had looked insignificant in that model's results except for the retired and student levels of job. The inclusion of age squared likely did a better job of predicting these levels, which are associated with particular ages.

The logit function is the natural log of odds, log(p/(1-p)), where p the probability of purchase and p/(1-p) defines the odds of purchase. Thus, the correct way of interpreting coefficients is to exponentiate them, providing the odds factor for a given predictor. The table below summarizes this interpretation for select features.

Feature	Coefficient	Interpretation
PC1	0.671	A one-unit change in PC1 results in increasing odds of purchase by exp(0.671)=196%. This better chance for purchases corresponds to lower employment, interest rates, and CPI.
Age and Age^2	-0.079 and 0.000926	A one-year increase in Age means applying a 92% factor due to age and a varying factor due to age squared. The varying factor has little effect at low ages and increasing effect as age rises. The overall impact of age and age squared is decreasing until age 43 and then increasing.
monthmar	0.970	The odds of purchase in March is 264% times that of the baseline month, April. Interpretations for other months are similar in this fashion. The highest month (where January and February are not available in the data) is March; the lowest month is May.

Task 9 – Investigate a shrinkage method (9 points)

Many candidates struggled to give a clear explanation of how elastic net performs feature selection, only giving some sense of formulas without noting what effect these formulas have.

Candidates sometimes used the wrong set of variables or did not comment on the differences in selected features as asked for.

Elastic net adds to the loglikelihood a penalty based on the magnitude of the estimated coefficients when training the model. The penalty includes both a term based on the sum of the squares of the coefficients, as in ridge regression, and a term based on the sum of the absolute values of the coefficients, as in LASSO regression. An alpha hyperparameter controls how much of each type of term is included, and a lambda hyperparameter controls the size of the overall penalty. The penalty induces shrinkage in the estimated coefficients when they are being optimized, and the inclusion of an absolute value term allows this shrinkage to go all the way to zero, effectively removing the feature.

I use elastic net with alpha equal to 0.5 to create a regularized regression, including the same variables (i.e., age, job, marital, edu_years, housing, loan, phone, month, weekday, and PC1) used in the full GLM model in Task 6. Using alpha equal to 0, which is ridge regression, is not appropriate for the goal of feature selection because the penalty based on solely on the squares of the coefficients cannot result in any of the coefficients shrinking all the way to zero, and therefore will not eliminate any variables.

Below is the output for the final elastic net regression model using the value of lambda that resulted in the minimum misclassification error:

```
35 x 1 sparse Matrix of class "dgCMatrix"
                          s0
(Intercept)       .
age               .
I(age^2)          .
jobblue-collar   -0.04168505
jobentrepreneur   .
jobhousemaid      .
jobmanagement     .
jobretired        0.18860248
jobself-employed  .
jobservices       .
jobstudent        0.12602323
jobtechnician     .
jobunemployed     .
jobunknown        .
maritalmarried    .
maritalsingle     .
edu_years         0.00387625
housingyes        .
housingunknown    .
loanyes           .
phonelandline     .
monthaug          .
monthdec          .
monthjul          0.05501910
monthjun          .
monthmar          0.44335928
monthmay         -0.69423107
monthnov         -0.10456389
monthoct          0.34932216
monthsep          0.30428194
weekdaymon        .
weekdaythu        .
weekdaytue        .
weekdaywed        .
PC1               0.54512418
```

The AUC were 0.7744 and 0.7828 on the training and test sets respectively.

The elastic net model includes factors from the following features:

- PC1
- Month (only March, May, July, September, October, November vs. all other months)
- Job (Only blue-collar, retired, and student vs. all other jobs)
- Edu_years

Edu_years, month, and PC1 also appeared in the Task 7 model, but the two age-based variables, loan, and phone were not selected by the elastic net model. However, due to binarization, not all categories within Month are given distinct coefficients as they had been given in the GLM in Task 7. The elastic net model did also select some Job levels, not chosen in the Task 7 model, including the most distinctive jobs by age. In choosing between using Job or the age-based variables, stepwise regression could only consider all or no levels of Job and found a better fit with the age-based variables, but elastic net could

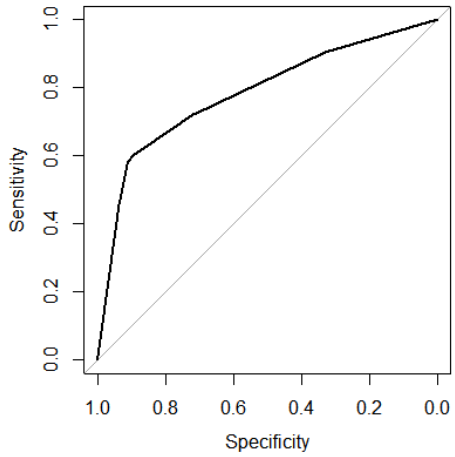
consider each level of Job distinctly and found particular levels provided a better fit, given the penalty, than the age-based variables.

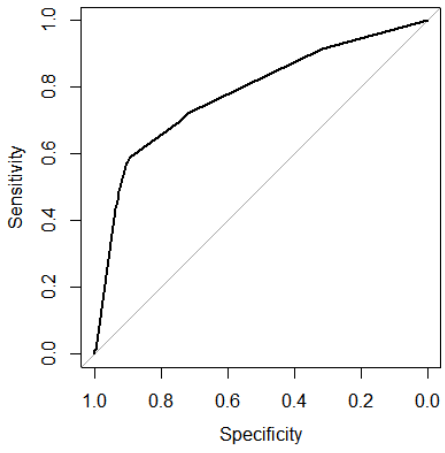
Task 10 – Construct a decision tree (7 points)

Candidates needed to have considerations related to the use of a decision tree when justifying the variables to be used.

Using underlying variables instead of the principal components derived from them is helpful here because it is harder to interpret those PCA variables and decision trees are not adversely affected by the presence of highly correlated variables. However, there is no or little information can be gained from splitting on employment after having split on irate, given that "irate" and "employment" are very highly correlated, and the selection of one or the other may be inconsistent depending on the selection of the training data. Therefore, dropping the employment variable is a reasonable choice.

Adding the square of age produces the exact same tree. For example, splitting on age > 30 versus age <= 30 is exactly the same as splitting on age squared > 900 versus age squared <= 900.





The ROC curves for the train (top) and test (bottom) data are shown above, with fairly similar shapes. As the AUC measures the area under the ROC curve, a similar story is evident when comparing the train

AUC of 0.7843 to the test AUC of 0.7827. In general, the higher probabilities of purchase predicted on the test data by the model fitted on the train data correspond to higher purchase rates, and lower probabilities correspond to lower purchase rates. Tiny exceptions can be seen in the ROC curve where the convexity is briefly positive, as in the bottom left—the very highest predicted probabilities do not correspond to the very highest purchase rates.

Task 11 – Employ cost-complexity pruning to construct a smaller tree (9 points)

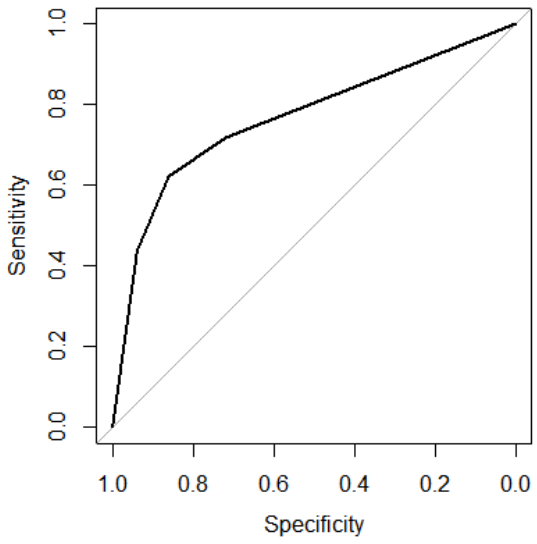
Better candidates explained what the cp table was doing rather than just noting mechanically how to choose the best cp value for pruning. Other valid techniques for pruning the tree than that shown were acceptable.

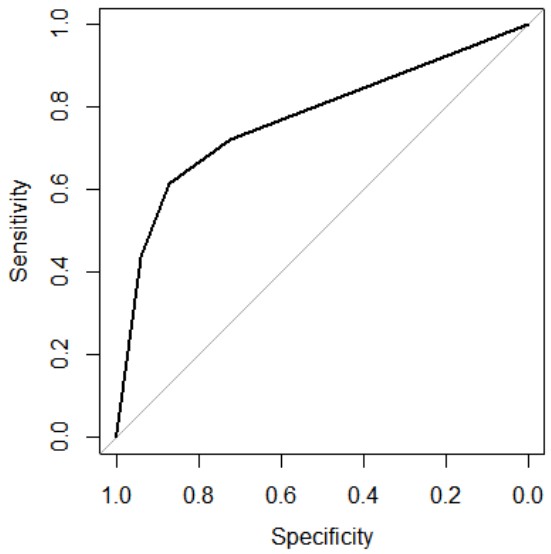
Better candidates pointed out the reduction in overfitting and interpreted the tree appropriately for marketing rather than just reading out the tree.

The complexity parameter (CP) is used to find the optimal tree size and reduce the overfitting seen above. The following output is from the initial unpruned tree. The optimal CP is the one that minimizes the cross validation error (in the xerror column). Row 6 accomplishes that, with CP = 0.0009466709. Pruning with this CP value will result in a tree with 8 splits and so 9 leaves.

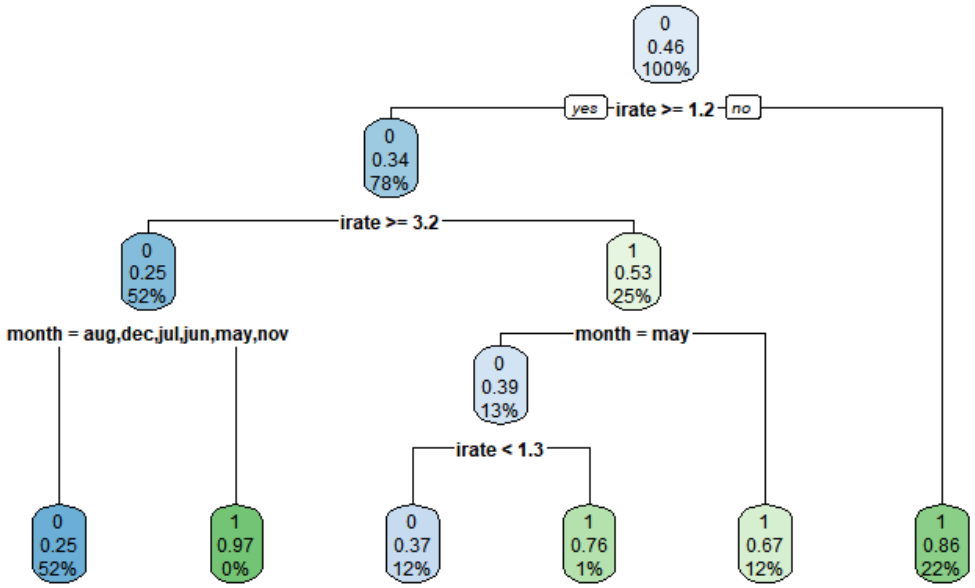
```
            CP nsplit rel error    xerror       xstd
1 0.3521615652      0 1.0000000 1.0000000 0.01306332
2 0.0448090880      1 0.6478384 0.6522562 0.01200730
3 0.0088355948      3 0.5582203 0.5629536 0.01147696
4 0.0085200379      4 0.5493847 0.5604292 0.01146015
5 0.0058903965      5 0.5408646 0.5389713 0.01131292
6 0.0009466709      8 0.5231934 0.5260334 0.01122035
7 0.0006311139      9 0.5222468 0.5298201 0.01124775
8 0.0005000000     12 0.5203534 0.5355002 0.01128837
```

For a tree with eight or less leaves, a complexity parameter of 0.006 is used to produce a six-leaf tree.





The train (upper) and test (lower) ROC curves are shown above. The test ROC curve no longer has areas with positive convexity, showing that overfitted predictions in those specific cases have been pruned. The train and test AUC's are 0.7688 and 0.7712 respectively. While the train AUC had to come down from the 0.7843 of the previous tree due to being a simpler model, the test AUC decreased by a smaller amount. The pruned test AUC is higher than the pruned train AUC, signaling that no overfitting is occurring with the simpler model and its predictors will be more reliable.



Of interest are the two leaves that account for the largest proportions of the training data.

- 52% of the past experience used for training falls into the leaf farthest to the left, when interest rates exceed 3.16% in the months from May to August or November to December. Only 25% of

these made a purchase, and the model predicts no purchase for future prospects in this situation. The combination of high interest rates and summer or holiday months appear to be a particularly poor combination for marketing our products, as this group had the lowest historical purchase rate of all eight groups.

- Another 22% of the past experience used for training falls into the leaf farthest to the right, when interest rates are less than 1.24%. 86% of these made a purchase, regardless of the month, and the model predicts a purchase for future prospects in this situation. Periods with very low interest rates are a good time to market our products regardless of other conditions.

Task 12 – Choose a model (4 points)

Some candidates did not consider both predictive power and applicability to the business problem, and others gave justifications based on one of these but then chose a model based on the other. This particular business problem did not favor choosing a model solely on AUC given how similar these typically were among models.

Model	Test AUC	Applicability
Full GLM (task 6)	0.7830	Low: shows trends, but no selection of variables
GLM StepAIC (task 8)	0.7842	Medium: Has selection, PC1, age^2 hard to explain
Elastic Net GLM (task 9)	0.7828	Medium-High: Has selection, PC1 hard to explain
Pruned Tree (task 11)	0.7712	Medium-High: easy to explain, very few variables

To choose a model I will use for advising the marketing department in future campaigns, I consider the balance between predictive power, indicated by a high value of test AUC, and a simpler model for giving more straightforward advice. The table above provides the test AUC's and some considerations for applicability.

I recommend using the elastic net GLM from Task 9. Among the three GLM models, the full GLM is worse on both test AUC and applicability, eliminating it from consideration. Of the other two, the one from the stepwise process is slightly more predictive but the one from elastic net is easier to explain as particular jobs are used instead of age and age squared to capture this source of variation. Given how close the accuracy is, the easier explanation to marketing is preferable. The pruned decision tree would be yet easier to explain, without having to deal with the principal component variable, but it has poorer accuracy and so few variables involved (just interest rate and month) that the insights may not be as useful to marketing as a more robust model would be. The elastic net GLM provides a good balance of accuracy and explainable insight generation.

Task 13 – Executive summary (20 points)

Rather than restating information from prior tasks, candidates were expected to alter their messaging for the intended audience. Often this includes avoiding overly technical language, discussing topics at a different level of detail, and translating performance metrics to be more meaningful to the reader. Brief discussions about approaches attempted are acceptable, but candidates should avoid lengthy discussion about models or techniques that were not ultimately

selected. The best candidates were able to incorporate the business context of the problem throughout their summary.

I have been asked to advise the marketing department at ABC insurance on what efforts will be most productive in terms of purchases for future marketing campaigns for a particular insurance product, based on data collected from a completed marketing campaign. The data has been analyzed using predictive models to bring out what aspects of the marketing campaign have the greatest impacts on whether customers purchase the product. Because the data is specific to just this product, this advice is also specific to just this product. The data did not include any experience in January or February, so no predictions on marketing campaigns for this product in these months can be made.

The data contains 10,000 observations for 14 variables that include personal information about the potential purchaser, the timing of the call, economic indicators at the time of the call, and whether a purchase was made, the variable to be predicted in the future based on the other variables. In this data, 46% of calls resulted in a purchase.

Some records had missing data. Almost 5% of the records did not have the education of the potential purchaser, and because these generally had higher purchase rates, I did not want to remove these records in case other variables helped to explain the higher purchase rates. The average education of other potential purchasers was used as a substitute, but I would like to discuss what leads to missing education data and whether this substitution is appropriate. Where other missing data was encountered, 142 records where dropped where this seemed insignificant to results but in other cases an "unknown" category was created. Modeling proceeded with the remaining 9,858 records.

Some of the economic indicators were highly correlated, so I applied a variety of techniques, from principal components analysis to simply removing one of the variables, employment, to improve the stability of our models.

To test the predictive power of all types of models, the models were trained on 70% of the data and their performance measured only on the remaining 30% of the data not yet seen by the models. The performance metric was "area under the curve", which describes how accurate the ranking of the probability of purchase is from highest to lowest when compared with purchases made.

During the modeling stage, I examined several options, including a full generalized linear model (GLM) and two reduced versions of this model to isolate the most important factors for distinguishing higher probabilities of purchase. All models in this stage performed reasonably well, but some are simpler than others. I also used a decision tree as it can handle interactions among factors more easily than GLMs, and after it was adjusted to retain only its most significant distinctions among potential purchasers, its results were somewhat less accurate than the GLM models. The decision tree only provided insights on interest rates and month where the best GLM model was more accurate and provided more insights, so the favored GLM model, called an elastic net model, was chosen for generating marketing insights.

A summary of this GLM model is provided below:

- Start by assuming every call has a 48.2% chance of resulting in a purchase. This is the same as an odds ratio of 0.93. The next steps will increase or decrease this odds ratio through multiplying by numbers higher or lower than 1, resulting in higher or lower probabilities of a purchase.
- Multiply by the odds ratio by the following odds factors for the job of who is being called:

    - Retired: 1.21 (highest purchase rates)
    - Student: 1.13 (higher purchase rates)
    - Other jobs: 1.00 (no adjustment)
    - Blue-collar: 0.96 (lower purchase rates)
- Then multiply the result by the following odds factors for education, a slight adjustment with more education leading to higher purchase rates (not all shown):
    - University degree: 1.06 (highest adjustment)
    - High school: 1.05
    - 6 years of education: 1.02
- Then multiply this result by the following odds factors for the month of the call:
    - March: 1.56 (highest purchase rates)
    - October: 1.42 (higher purchase rates)
    - September: 1.36
    - July: 1.06
    - Other months: 1.00 (no adjustment)
    - November: 0.90 (lower purchase rates)
    - May: 0.50 (lowest purchase rates)
- Then multiply this result by a "economic factor" that takes into account how different a set of economic indicators are from their average values observed during the last marketing campaign. In general, purchase rates increase when interest rates, employment at ABC, and the consumer price index move down.
- Convert back to a probability by taking this result and dividing by one plus itself.

I will provide a spreadsheet that carries out this calculation but provide the details above to specify which situations are expected to have positive or negative impacts on purchase rates.

The most impactful factors are about timing: month and interest rates/employment, the latter being pair of factors often moving in tandem. The highest purchase rates, everything else being equal, occur in March, September, and October, while particularly low purchase rates occur in May. In the prior marketing campaign, far more calls were made in May than any other month, so I wonder whether the volume of calls itself was a factor in the proportion of purchases—this should be studied further. On interest rates/employment, calls will be more productive when these are lower, with another model indicating that particular low interest rates, under 1.24% lead to high purchase rates no matter the month.

Smaller impacts on call success are about who is called: job and education level. I recommend targeting retired and student prospects above all others, with a slight nudge away from those with blue-collar jobs. Education level has only a small impact, but targeting those with higher levels of education may produce slightly higher purchase rates.

At a high level, a successful marketing campaign for this product has more to do with market conditions and timing of the calls and less to do with the characteristics of who is called. This conclusion is dependent on the data provided and techniques used. I look forward to discussing these results with you in more detail and working together to refine the insights generated thus far.