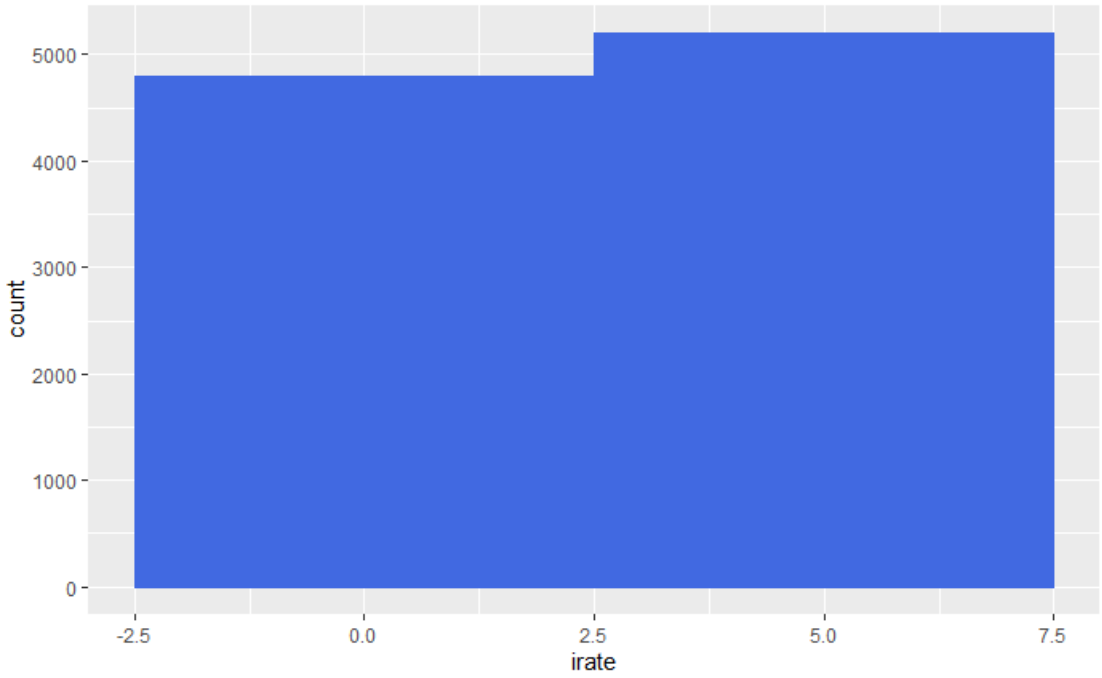
Practice Exam – Customer Phone Calls (SOA PA 6/18/20) Solution

Task 1 – Explore the data (8 points)

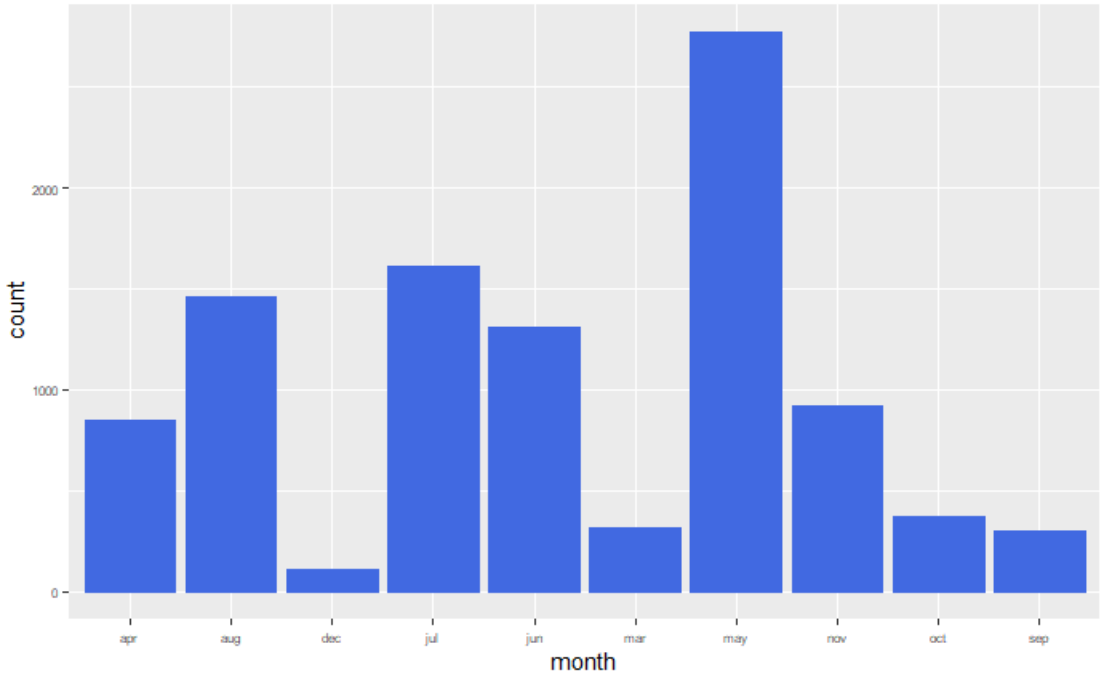
Irate



The bar plot above shows short-term interest rates. This graph is not informative, but was created by my assistant. You can see from the summary below that the range of irates go from 0.634 to 5.045. This may be predictive of whether a customer purchases the product, but further information on purchase percentage would be needed in order to make this determination.

```
      age            job              marital            housing              loan          
 Min.   :17.00   Length:10000       Length:10000       Length:10000       Length:10000      
 1st Qu.:32.00   Class :character   Class :character   Class :character   Class :character  
 Median :38.00   Mode  :character   Mode  :character   Mode  :character   Mode  :character  
 Mean   :40.44                                                                              
 3rd Qu.:48.00                                                                              
 Max.   :98.00                                                                              

     phone              month              weekday               CPI             CCI        
 Length:10000       Length:10000       Length:10000       Min.   :92.20   Min.   :-50.80  
 Class :character   Class :character   Class :character   1st Qu.:92.89   1st Qu.:-42.70  
 Mode  :character   Mode  :character   Mode  :character   Median :93.44   Median :-41.80  
                                                          Mean   :93.49   Mean   :-40.26  
                                                          3rd Qu.:93.99   3rd Qu.:-36.40  
                                                          Max.   :94.77   Max.   :-26.90  

     irate          employment      purchase       edu_years     
 Min.   :0.634   Min.   :4964   Min.   :0.000   Min.   : 1.00  
 1st Qu.:1.250   1st Qu.:5076   1st Qu.:0.000   1st Qu.: 9.00  
 Median :4.076   Median :5191   Median :0.000   Median :12.00  
 Mean   :3.030   Mean   :5139   Mean   :0.464   Mean   :11.92  
 3rd Qu.:4.959   3rd Qu.:5228   3rd Qu.:1.000   3rd Qu.:16.00  
 Max.   :5.045   Max.   :5228   Max.   :1.000   Max.   :16.00  
                                                NA's   :468    
```

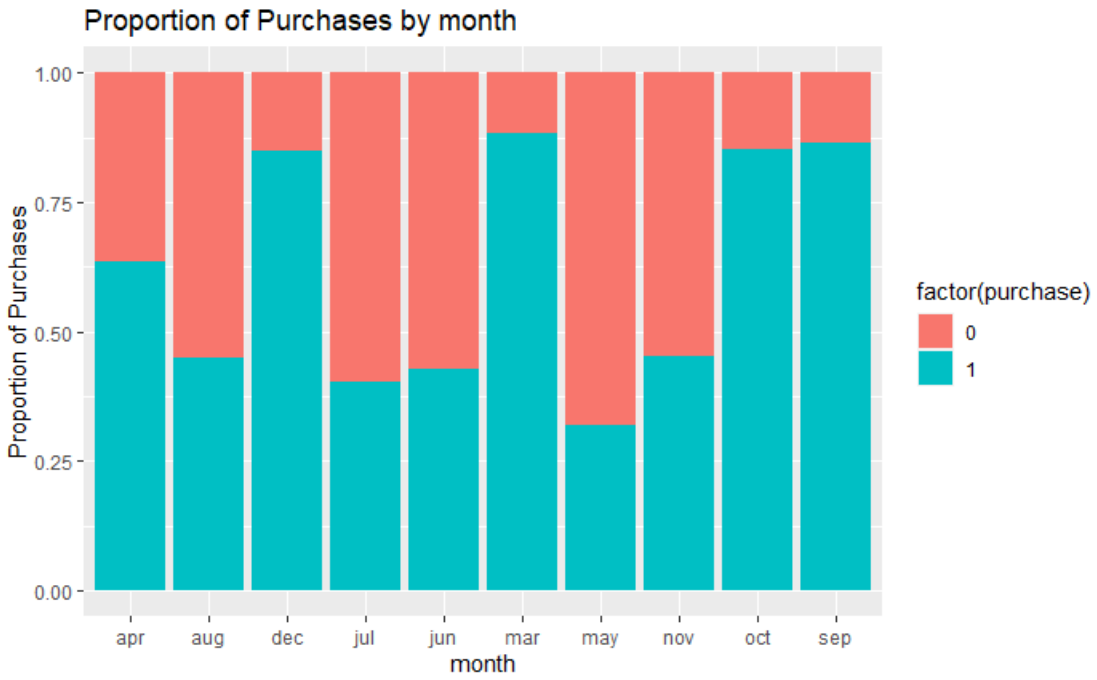
Month



The above bar chart shows the number of customer calls (y-axis) for each month of the year (x-axis). The months of August, July, and May have the highest call volumes. This may be because customers are more active during these times. ABC Insurance could use this data to justify increasing their staff levels during these months to meet customer demand and obtain additional business.

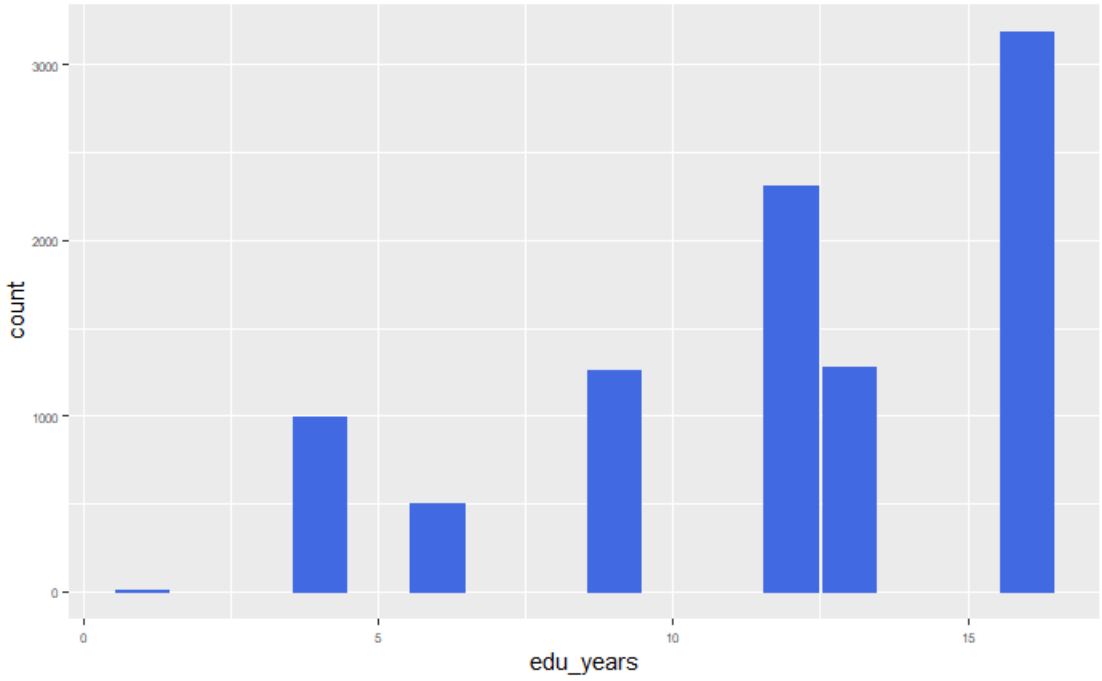
Note: These are shown in alphabetical order (instead of calendar order) because R orders factors in this way.

It is also important to determine if purchases of insurance are seasonal. The graph below shows the months in which marketing calls were made (x-axis) and the purchase percentage (y-axis) shown by the blue bars. The months of highest volume are December, March, October, and September. The marketing team should increase calls during these months to take advantage of this increase in demand. They could also increase working hours and employ more sales representatives.



When comparing the above chart to the prior chart, we notice a relationship between the percentage of customers who purchase and the number of customers in that month. The months with the most observations, August, July, and May, also have the lowest purchase rates. Months with few observations, December, March, and September, have the highest purchase rates.

Edu_years

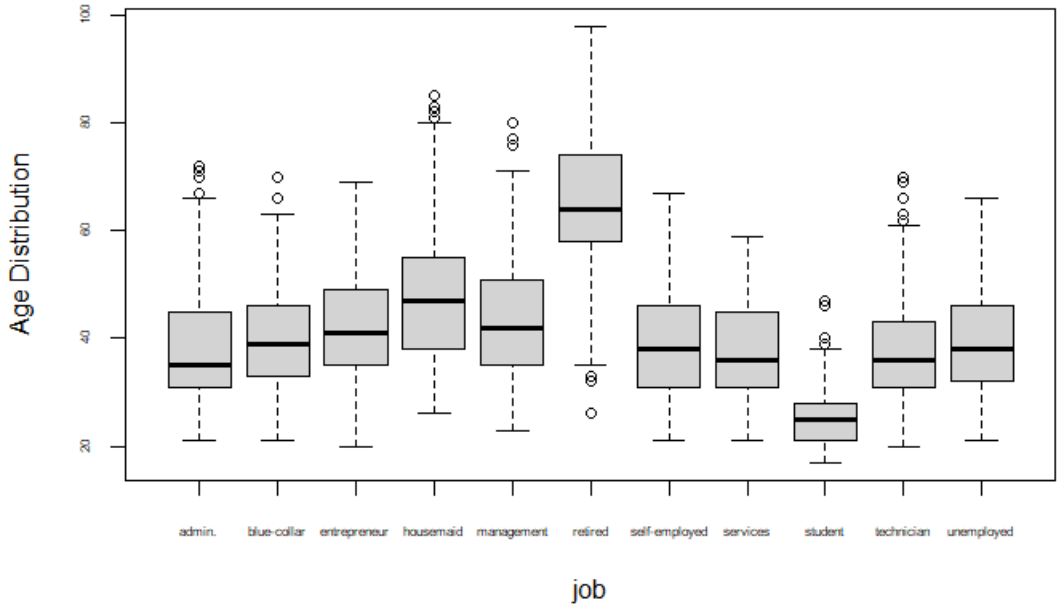


The above graph shows the prospective customers' education levels (x-axis) and the number of customers (y-axis). Most of ABC's customers are well-educated. This graph would be easier to read if there were labels on the x-axis. Most customers have 16 years of education which means that they have obtained an undergraduate degree. ABC Insurance may be able to directly market to graduating college seniors as they are more likely to purchase than those who have not completed college or are still in school. There are 468 missing values, and these are not explained. In addition, there are 5 customers who each had only 1 year of education. This seems unlikely and is therefore an issue that ABC should follow up on.

edu_years <int>	n <int>
1	5
4	997
6	499
9	1261
12	2305
13	1280
16	3185
NA	468

8 rows

Job

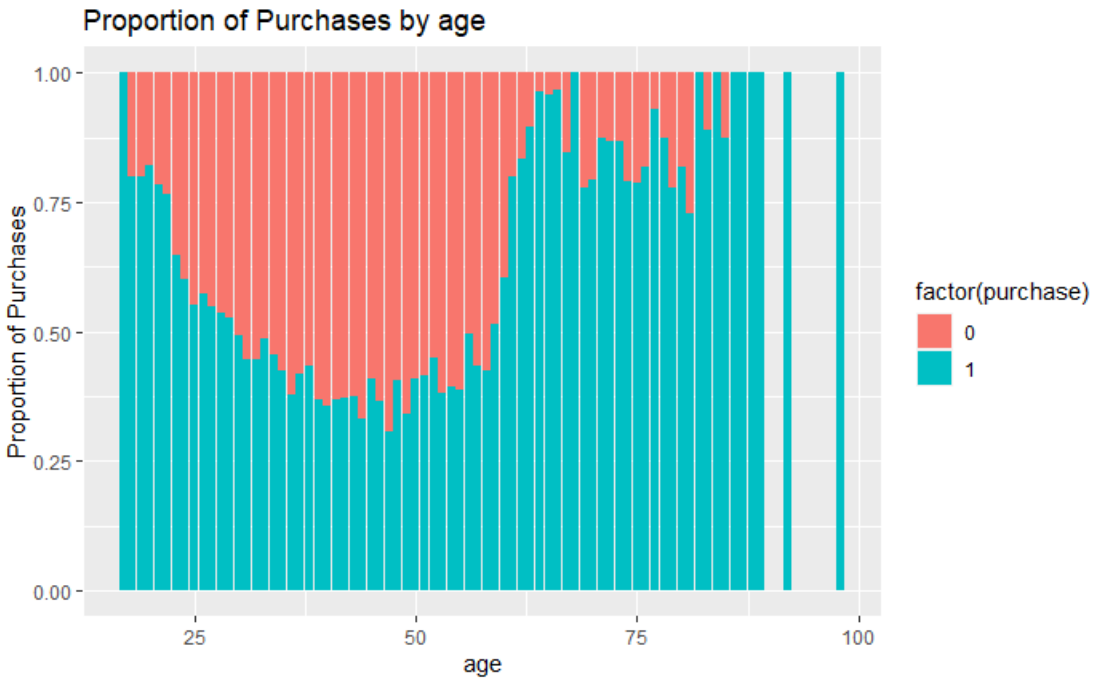


The boxplot above shows the customer job categories (x-axis) and age distribution (y-axis). These box plots show median ages as the center bars, the upper 75th percentiles as the top bars, and the 25th percentiles as the lower bars. ABC should customize their marketing strategies according to potential customers' ages and occupations. For example:

- Customers who are retired are likely older and may be more comfortable speaking on the phone than using text services such as chats or email;
- Customers who are students may prefer text communication and be more responsive to social media advertising;
- "Technicians" tend to be between 25 and 45 years old and therefore may not have a clear preference for communication type;
- Workers between 25-50 years old, including blue-color workers, entrepreneurs, and housemaids, as well as those who are self-employed, in management, or work in the service sector might all be reached through similar age-appropriate marketing strategies. For example, younger people would probably be less interested in retirement products than older people.

Age

The graph shows that current age (x-axis) strongly influences the decision to purchase insurance (y-axis). People in their twenties purchase insurance about 75% of the time, but those in their thirties purchase only about 50% of the time. Upon reaching the retirement age of 65, the purchase percentage increases dramatically to 80%. This is a non-linear relationship, and is logical because older people may be in poorer health and benefit more from having insurance.

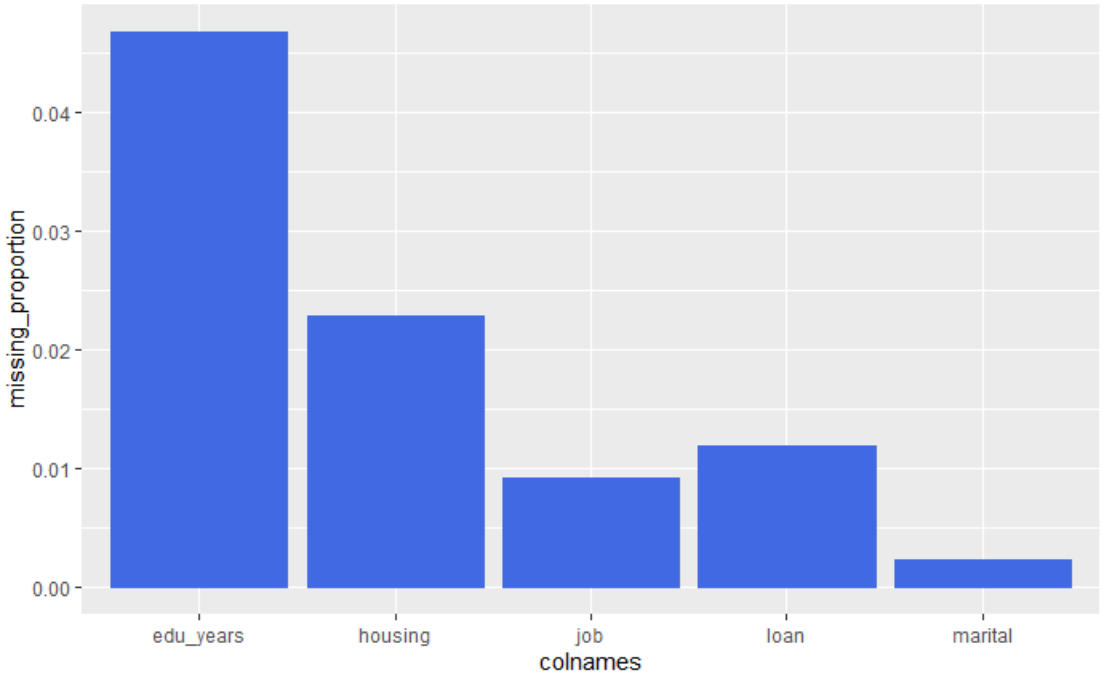


Task 2 – Consider the education variable (3 points)

If edu_years is considered a factor it would have 7 categories, requiring a model to include 6 dummy values or indicator columns. However, if it is treated as numeric it only requires one column. The advantage to using a numeric value is that it reduces the dimensions of the data and helps to combat the curse of dimensionality, stating that as the number of predictors increases, the density of the feature space decreases and so more observations are needed.

Regardless of whether edu_years is a factor or numeric, the decision tree can fit non-linear splits and the GLM cannot. This means that a tree is more flexible. When using a numeric variable, the GLM has only 1 coefficient for this predictor, which results in a simpler model. The factor would result in 6 coefficients. If using a tree, the complexity would be about the same regardless of the type. But choosing numeric makes the rules easier to interpret because edu_years has a high cardinality.

Task 3 – Handle missing values (5 points)



Edu_years

The category edu_years has 468 (about 5%) of the values missing. They might have been omitted because the subjects never completed their education; however, it is more likely that these are errors. This omission could be either handled by either removing these people from the data or replacing the missing values with an average number of years. People whose records do not include an edu_years value have a purchase percentage of 54% whereas people who have a complete record have a lower purchase percentage (46%). This implies that the missing records might be significant, and therefore should not be removed. Therefore, I choose to include the records by imputing the mean.

Housing

Housing is missing in about 2.5% of cases. As the difference in purchase percentage is small (47% vs 46%) I chose to remove these values. The cause of the omission is likely random and so removing them will not deprive the marketing team of important information regarding likelihood of purchase.

Job

Employment Information is missing in only about 1% of cases. There might be significance here as the purchase percentage for those with a missing job field is 41% vs. 46% for those who have complete records. It could be that omissions in this field reflect unemployment, and that people with unlisted jobs are at a higher financial risk and thus more likely to purchase insurance. For this reason, I indicate the values as “missing”.

Loan

There were no missing values in the loan category after removal of the records which had missing job data. The purchase percentage was 45% for both missing and non-missing cases, so removal does not result in a loss of predictive power.

marital

After removing the cases with incomplete data for loans and housing, 23 cases remained with a missing value for marital status. These cases show a significant difference in the purchase percentage with complete records at 46% and incomplete at 52%. However, due to the low number of cases, I felt that it was justifiable to remove these values as this action would not have an important effect.

Task 4 – Investigate correlations (3 points)

I compare the correlations of the linear variables below. Two variables are correlated when there is a tendency of one variable to increase as the other variable increases and to decrease when that variable decreases.

- edu_years and age have a negative correlation;
- Interest rate and CPI have a very high correlation (0.59); and
- Employment and CPI have a high correlation (0.98).

	age	edu_years	CPI	CCI	irate	employme
age	1.00	-0.24	-0.02	0.14	-0.04	-0.07
edu_years	-0.24	1.00	-0.09	0.04	-0.08	-0.08
CPI	-0.02	-0.09	1.00	-0.14	0.59	0.38
CCI	0.14	0.04	-0.14	1.00	0.06	-0.08
irate	-0.04	-0.08	0.59	0.06	1.00	0.94
employment	-0.07	-0.08	0.38	-0.08	0.94	1.00

One of the assumptions of GLMs is that the predictor variables are not correlated. Correlations between irate and employment may cause problems when fitting the model. Decision trees, on the other hand, handle correlations well and would not be impacted.

GLMs perform poorly because the separate coefficients for each of the correlated variables have an offsetting impact. This results in an unstable model with coefficients that can change drastically depending on the training split. The p-values for the correlated variables would likely be large.

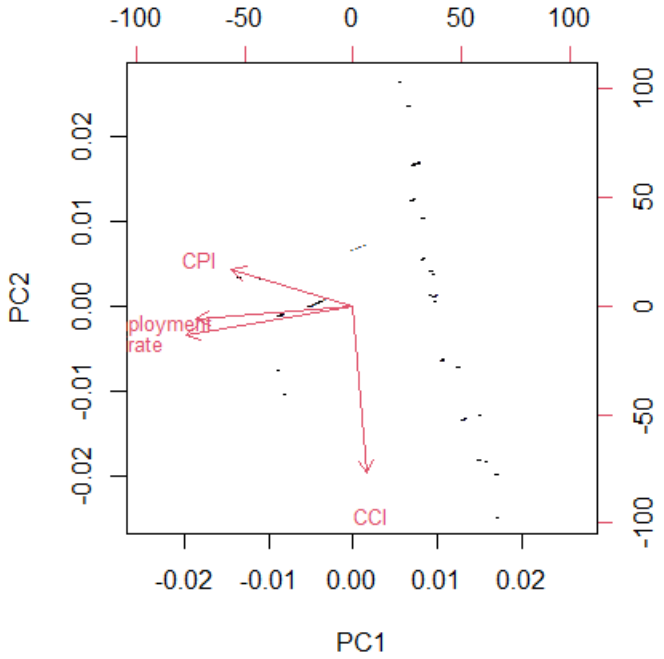
Trees are not strongly impacted by this because only one variable is selected at each split. In other words, the feature selection attribute of trees will choose only one of the correlated variables at a time.

1) Remove one of the correlated variables and leave the others, eliminating the issue of having counteracting coefficients and an unstable model.
2) Use a stepwise selection such as Step AIC or Step BIC to remove a subset of the correlated variables automatically. Once one variable is included in the model, adding another correlated variable would not improve the AIC or BIC and so this model variation would not be used.
3) Use lasso regression to remove variables which do not decrease the test error (root mean squared error) based on cross-validation.
4) Use clustering to create a new feature which captures the patterns. Customers who have similar cluster characteristics for correlated variables would be grouped together. For instance, k-means can be used on the economic variables.

[Note: only one answer is needed. Four are provided here to offer additional options.]

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

Principal component analysis (PCA) is a dimensionality reduction method which simplifies the original variables. PCA is a good way to handle the economic variable irate, cci, cpi, and employment as earlier analysis showed that these are correlated. PCA begins with the original variables and then applies scaling by subtracting the mean and dividing by the standard deviation. It then creates linear combinations which are uncorrelated (orthogonal).



The above biplot shows projections of the original variables onto the first two principal components. CPI, employment, and irate are grouped together because they are correlated. The nearly perpendicular direction of CPI is logical because the correlation between CCI and irate is only 0.06.

```
Rotation (n x k) = (4 x 4):
                    PC1         PC2        PC3        PC4
CPI         -0.46880713  0.21391085  0.8292006  0.2165374
CCI          0.05532036 -0.95955084  0.2462228  0.1248046
irate       -0.64350767 -0.16588933 -0.1288116 -0.7360613
employment  -0.60254246 -0.07736314 -0.4849828  0.6290860
```

The recipe for the first principal component follows. This can be interpreted as an economic indicator - high when the economy is in a recession with low interest rates, employment, and prices (CPI) and low when the economy is growing.

-(0.47)(CPI) + (0.06)(CCI) – (0.64)(irate) – (0.6)(employment)

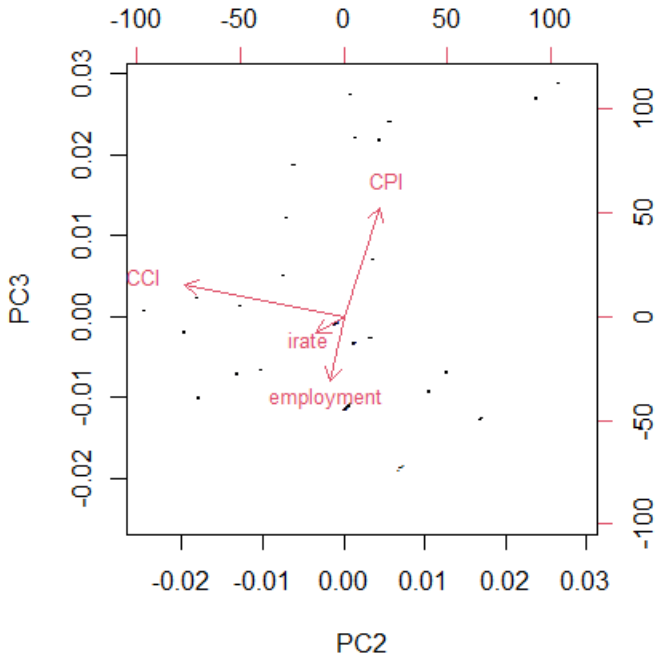
The recipe for the second follows. This PC has a larger rotation on CCI (0.96) than the first PC.

(0.21)(CPI) – (0.96)(CCI) – (0.17)(irate) – (0.08)(employment)

As age and or edu_years are unrelated to the state of the economy and have an exceptionally low correlation, they have not been added to this PCA. Adding them would have made the PCA more difficult to interpret and the variables would not have explained any differences in irate, cpi, cci and/or employment.

(Optional) - You would not be expected to do this because it requires changing the code, but you can also look at the biplot of the second and third PCs using:

biplot(pca, **choices = 2:3**,cex = 0.8, xlabs = rep(".", nrow(tmp)))



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

First, the data was divided into training (70%) and test (30%). The model is trained using the training data and the performance is evaluated based on the test data. I verified that the purchase percentage was the same (about 46%) in each data set.

I fit a GLM that used only age as a predictor.

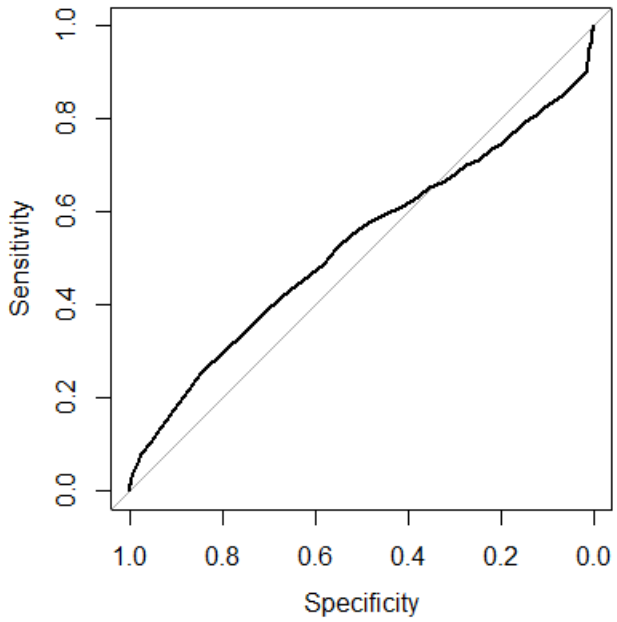
```
Call:
glm(formula = purchase ~ age, family = binomial(link = "logit"),
    data = data_train)

Deviance Residuals:
   Min      1Q  Median      3Q     Max
-1.267  -1.115  -1.077   1.244   1.315
```

```
Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept) -0.450080   0.086333  -5.213 1.85e-07 ***
age          0.007737   0.002050   3.773 0.000161 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 9429.2  on 6824  degrees of freedom
Residual deviance: 9414.9  on 6823  degrees of freedom
AIC: 9418.9
```

Then I first a second GLM using a binomial family and a logit link that included all variables except for the economic ones. These were instead included in the principal component.

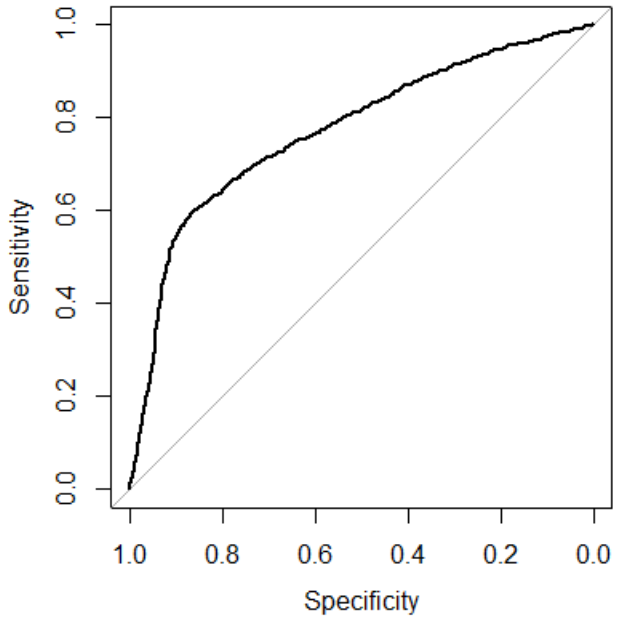
```
Call:
glm(formula = purchase ~ age + job + marital + edu_years + housing +
    loan + phone + month + weekday + PC1, family = binomial(link = "logit"),
    data = data_train)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.5973  -0.8632  -0.5438   0.8517   2.1627

Coefficients:
                 Estimate Std. Error z value Pr(>|z|)
(Intercept)      -0.631872   0.260079  -2.430  0.01512 *
age               0.002946   0.003384   0.871  0.38394
jobblue-collar   -0.114017   0.101421  -1.124  0.26093
jobentrepreneur  -0.236706   0.167864  -1.410  0.15851
jobhousemaid     -0.100514   0.200305  -0.502  0.61580
jobmanagement    -0.009585   0.119451  -0.080  0.93604
jobMissing       -0.250239   0.321181  -0.779  0.43591
jobretired        0.374440   0.161834   2.314  0.02068 *
jobself-employed  0.036869   0.164861   0.224  0.82304
jobservices      -0.121575   0.109960  -1.106  0.26889
jobstudent        0.472497   0.188618   2.505  0.01224 *
jobtechnician    -0.063646   0.088016  -0.723  0.46961
jobunemployed     0.090171   0.179249   0.503  0.61493
maritalmarried    0.027503   0.093066   0.296  0.76760
maritalsingle     0.026463   0.106707   0.248  0.80414
edu_years         0.022373   0.009823   2.278  0.02275 *
housingyes       -0.030172   0.056918  -0.530  0.59604
loanyes          -0.073835   0.077677  -0.951  0.34184
phonelandline     0.092480   0.089770   1.030  0.30292
monthaug          0.213264   0.123519   1.727  0.08424 .
monthdec          0.674422   0.377454   1.787  0.07397 .
monthjul          0.645912   0.128755   5.017 5.26e-07 ***
monthjun          0.462138   0.129784   3.561  0.00037 ***
monthmar          1.076071   0.222971   4.826 1.39e-06 ***
monthmay         -0.590032   0.105687  -5.583 2.37e-08 ***
monthnov          0.024638   0.128588   0.192  0.84805
monthoct          1.206390   0.223308   5.402 6.58e-08 ***
monthsep          0.638003   0.222432   2.868  0.00413 **
weekdaymon       -0.255756   0.091312  -2.801  0.00510 **
weekdaythu        0.057448   0.088828   0.647  0.51781
weekdaytue        0.036495   0.092138   0.396  0.69204
weekdaywed        0.072533   0.090667   0.800  0.42371
PC1               0.695169   0.029230  23.783  < 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: 9429.2  on 6824  degrees of freedom
Residual deviance: 7562.2  on 6792  degrees of freedom
AIC: 7628.2
```




```
Area under the curve: 0.5271
```

The results of this model are not very predictive. An AUC of 0.5 reflects random guessing.



```
Area under the curve: 0.7714
```

The first model has a positive sign on age; the purchase percentage increases as age increases. The p-value is smaller, indicating that it is adding predictive value to the model.

The second GLM, which uses all other variables, also has a positive sign on age. However, the p-value here is large (0.38) meaning that once all other variables are considered, age is no longer a predictor of insurance purchases. This finding is logical because knowing a person's housing status, job, loan history, and educational history along with the month of the year and other information is sufficient to predict purchases without considering age.

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

In the data exploration task, I noticed that the purchase percentage by age starts out high, falls for people from 30 – 45 years old, and then increases again at age 65. Using age alone in a GLM would fail to capture this pattern. Using a single variable results in only one coefficient, so the impact can only be positive or negative. In a decision tree, this is not a problem because the separate nodes are able to predict different purchase percentages. By adding in a square of age, I allow for the linear predictor to capture parabolic trends, including the variable pattern evident in the graph. This would likely improve the model.

The best subset selection algorithm needs to consider all possibilities, including models with just age, age and edu_years, age and edu_years and marital, age and edu_years and marital and CPI, etc. A search through all possible combinations can find the subset of variables with the best performance.

One advantage to using a stepwise forward or backward selection is that fewer models are considered, which reduces the likelihood of overfitting. Because so many models are considered in the most effective and efficient subset selections, it is possible that some random combinations have a good model score even though the relation may not reflect reality. Using a stepwise selection looks at fewer models and so is less likely to overfit and the identified model tends to perform better on new test data.

One disadvantage to using stepwise forward or backward selection is that fewer models are considered and so it is possible that a better performing model exists. Forward selection starts with no variables and only adds those which improve the scores. Backward selection starts with all of the variables and removes them until the best model is found. The choice of using either forward or backward selection is somewhat arbitrary and is determined by the actuary running the model.

Using backward selection will result in a simpler model that will be easy for the marketing department to understand. The resulting model can be explained in a spreadsheet using only a person's age, edu_years, the month, weekday, and the PC score. This eliminates the variables of job, marital, housing, loan, and phone. Using fewer variables is more cost efficient because data needs to be purchased, collected from customers, and stored in databases.

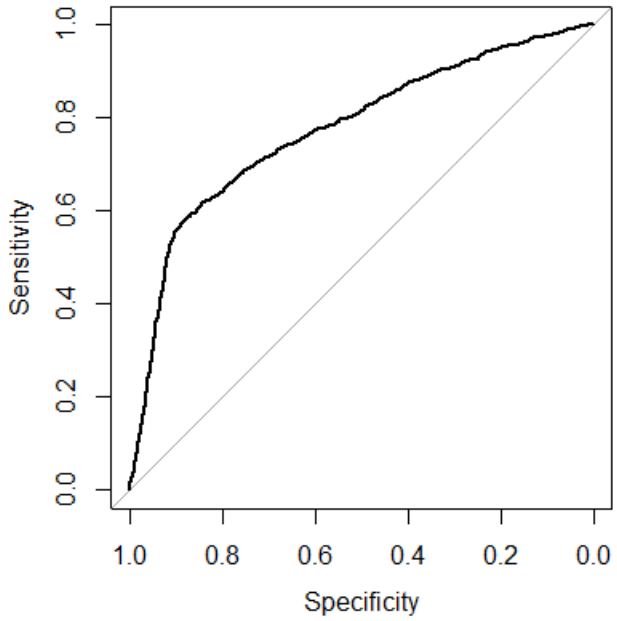
The remaining variables all have small p-values except for certain months and weekdays. The small p-value shows that the age squared predictor is adding predictive value. The signs of the coefficients

indicate if a variable increases or decreases the purchase percentage. The signs of age and age squared are opposite each other, consistent with the purchase percentage decreasing and then increasing.

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

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.8880  -0.8688  -0.5372   0.8693   2.1912

Coefficients:
              Estimate Std. Error z value Pr(>|z|)
(Intercept)  0.8775689  0.3664558   2.395 0.016632 *
age         -0.0755878  0.0159958  -4.725 2.30e-06 ***
I(age^2)     0.0009097  0.0001803   5.046 4.52e-07 ***
edu_years    0.0288556  0.0078855   3.659 0.000253 ***
monthaug     0.2019588  0.1198154   1.686 0.091876 .
monthdec     0.7122913  0.3769883   1.889 0.058835 .
monthjul     0.5790158  0.1242697   4.659 3.17e-06 ***
monthjun     0.4751246  0.1274522   3.728 0.000193 ***
monthmar     1.0615582  0.2232158   4.756 1.98e-06 ***
monthmay    -0.5771931  0.1042858  -5.535 3.12e-08 ***
monthnov     0.0162034  0.1270744   0.128 0.898536
monthoct     1.2305853  0.2223714   5.534 3.13e-08 ***
monthsep     0.6541006  0.2218739   2.948 0.003198 **
weekdaymon  -0.2688710  0.0910455  -2.953 0.003145 **
weekdaythu   0.0494822  0.0887055   0.558 0.576964
weekdaytue   0.0236469  0.0921108   0.257 0.797393
weekdaywed   0.0611707  0.0905478   0.676 0.499318
PC1          0.6682684  0.0245336  27.239  < 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: 9429.2  on 6824  degrees of freedom
Residual deviance: 7560.6  on 6807  degrees of freedom
AIC: 7596.6

Number of Fisher Scoring iterations: 4
```

Task 8 – Evaluate the model (12 points)

The AUC on the model looks good. It showed 0.7836 on training data and 0.774 on test data.



The AUC shows the probability that a randomly chosen purchaser is ranked higher than a randomly chosen non-purchaser. An AUC of 1.0 would be a perfect model. The ROC curve would be in the upper left-hand corner, showing that the sensitivity is 1.0 and specificity is 1.0 for all values of the cutoff. An AUC of 0.5 would indicate a model with just random guesses, with each customer given a 47% chance of being a 1 and 53% of being a 0. An AUC of less than 0.5 indicates that something is wrong, that either the model is overfitting so that the training AUC is greater than 0.5 but less than 0.5 on the test data, or that there is an error in the calculation.

In Task 1, I noticed that there were two months, December and March, which had few customers. These months have larger p-values because there is less confidence in the coefficient estimates. I noticed that there were some edu years which had no customers in them. This is acceptable because it is a numeric variable instead of a factor and the p-value is small. The age variable by itself was not predictive but adding age squared was able to account for the non-linear trend and so both were predictive. The coefficients have opposite signs indicating the change in direction.

The logistic regression can be interpreted so that a product of numbers creates log odds of a customer purchase. This is the probability that a customer purchases divided by the probability that they do not purchase. It starts with the reference level, Friday during the month of April.

The marketing department can use the following table to help them better understand their customers:

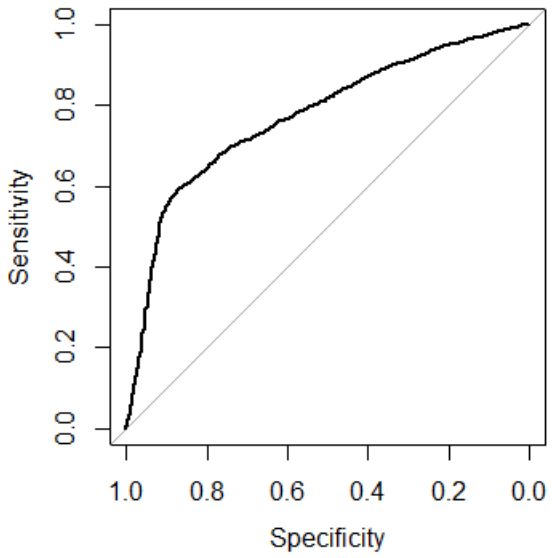
Variable	Interpretation
Age and Age Squared	The purchase log odds are highest for younger people and older people. The log odds decrease by a factor of 0.93 for each year of age and increases by a factor of 1.001 for each value of age squared.
Years of Education	The log odds increase as the education increases. Multiply by 1.029^n where n is the person's years of education.
Month of August	The log odds are 22.5% higher for the month of August as compared to the reference month of April.
Weekday Monday	The log odds are 23.6% lower on Mondays as compared to the reference day of Fridays
Principal Component	The logs odds increase by a factor of 1.951^n where n is the value of the principal component score. This is the economic score which increase as the interest rates, employment, and confidence index increase.

Task 9 – Investigate a shrinkage method (9 points)

Elastic net performs feature selection by maximizing an objective function. This includes the negative log likelihood as well as a second term. In the case of the LASSO, this is the sum of the absolute value of the coefficients, which removes variables by setting their coefficients to zero. A Ridge regression uses the sum of the squares of the coefficients, which makes the coefficients smaller. There is also a third case, a combination of the LASSO and Ridge known as the elastic net, which uses a weighted average of the first and second penalty terms. This is controlled by the alpha parameter. When alpha = 1, the model is the LASSO. When alpha = 0, the model is the Ridge.

If our goal is to remove variables, then we need to set alpha equal to 1. If we want to shrink the size of the coefficients, then we should choose an alpha greater than zero. Setting alpha = 0 is the ridge model which does not remove any features.

The AUC is 0.784 on the training data and 0.7729 on the test data.



This model uses the same variables as the GLM from Task 7. However, the coefficients are generally smaller in terms of absolute value. For age, the GLM has a coefficient of -0.08 but this is -0.05 in the elastic net. This is because the algorithm shrinks the coefficients. The signs are always the same.

```
(Intercept)       .
age              -0.0500231036
I(age^2)          0.0006101884
jobblue-collar   -0.0960492398
jobentrepreneur  -0.2173811278
jobhousemaid     -0.1169600585
jobmanagement    -0.0032929731
jobMissing       -0.2815233570
jobretired        0.1528720831
jobself-employed  0.0331764153
jobservices      -0.1158033136
jobstudent        0.3670960441
jobtechnician    -0.0522769075
jobunemployed     0.0946419011
maritalmarried    0.0348001759
maritalsingle    -0.0039754466
edu_years         0.0248261983
housingyes       -0.0281307300
loanyes          -0.0788513986
phonelandline     0.0774394599
monthaug          0.1930661003
monthdec          0.6612708573
monthjul          0.6020803680
monthjun          0.4359557175
monthmar          1.0274910521
monthmay         -0.5889368647
monthnov          0.0116038172
monthoct          1.1777909419
monthsep          0.6143461497
weekdaymon       -0.2600274984
weekdaythu        0.0504308674
weekdaytue        0.0280444118
weekdaywed        0.0617459256
PC1               0.6765241462
```

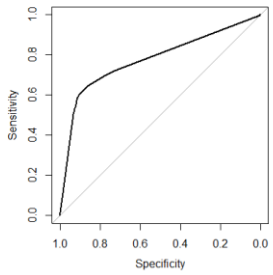
Task 10 – Construct a decision tree (7 points)

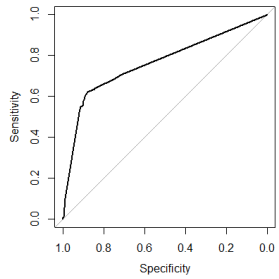
Decision trees perform variable selection automatically. This means that the correlated variables included in the principal component analysis can be used separately without causing problems in the model. For GLMs, including variables which have a high correlation leads to instability, but this is not an issue for tree-based models. Using these variables has the additional benefit of making a tree easier to interpret.

Because I want these results to be useful for the marketing team at ABC Insurance, I chose to remove employees and to use irate instead. This will help them to take interest rates into consideration as they assist their sales team in selling insurance products to customers.

Decision trees automatically capture non-linear relationships. For example, in the age variable the tree can use a cut points at 25, 40, and over 40 so that the predicted values are higher for customers under age 25, lower for those between ages 25 - 40, and higher for older customers. The age squared variable is only needed for the GLM, because it does not capture these non-linear trends.

This showed a good result with the training at 0.7783 and the test AUC at 0.7634. If the tree was overfitting, the AUC on the test data would be much lower than on the training data. However, these results are not as good as the elastic net, with its test AUC of 0.77.





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

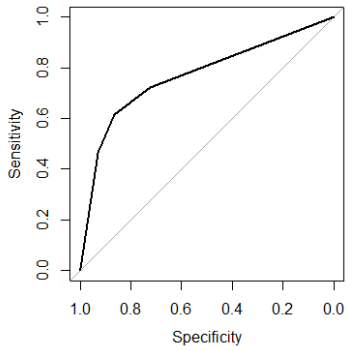
The table below shows the tree's complexity parameters and its number of branches along with a misclassification error, xerror. The cp parameter controls the height and variance of the tree. Any split

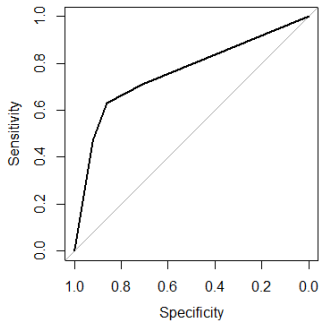
which does not decrease the overall lack of fit by a factor of cp is not used. There are 12 different tree fits, one for each row. The one which has the lowest error uses a CP of 0.00063, which has 14 splits.

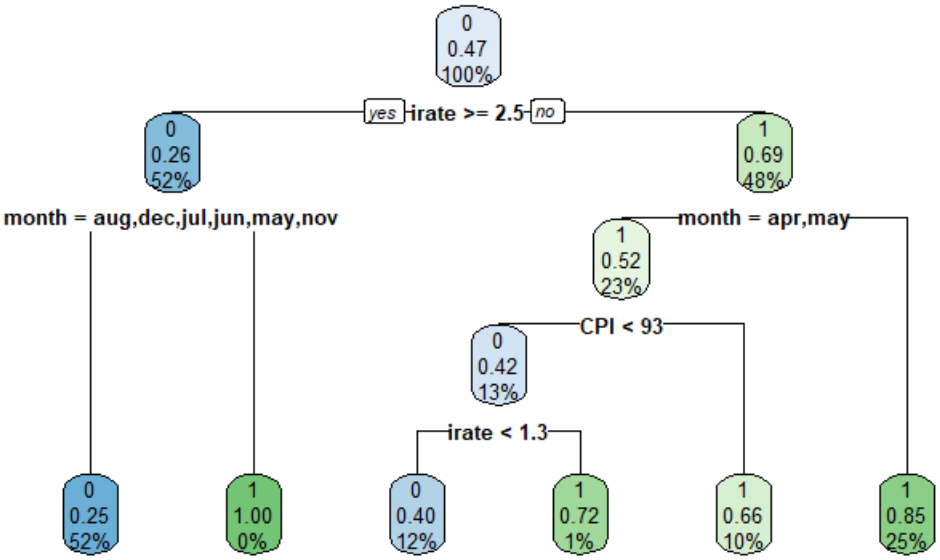
	CP	nsplit	rel error	xerror	xstd
1	0.3955317810	0	1.0000000	1.0000000	0.01296700
2	0.0228130900	1	0.6044682	0.6261800	0.01181461
3	0.0097545626	3	0.5588420	0.5657646	0.01145102
4	0.0069225928	4	0.5490875	0.5610447	0.01142015
5	0.0061359346	5	0.5421649	0.5610447	0.01142015
6	0.0037759597	7	0.5298930	0.5427942	0.01129729
7	0.0023599748	8	0.5261171	0.5355570	0.01124700
8	0.0022026432	11	0.5179358	0.5355570	0.01124700
9	0.0018879799	12	0.5157332	0.5377596	0.01126240
10	0.0015733166	13	0.5138452	0.5383889	0.01126679
11	**0.0006293266**	**14**	**0.5122719**	**0.5339836**	**0.01123595**
12	0.0005000000	21	0.5072373	0.5402769	0.01127990

It was important to find a tree that had 8 leaves, its terminal nodes. The table above shows the number of splits, the internal nodes. The CP parameter was increased to make the tree simpler and reduce the number of leaves. I found that setting CP to 0.00613 resulted in 6 leaves. Other higher values in the table all resulted in more than 8 leaves

The AUC is 0.7691 on the training data and 0.7613 on the test data.







The above tree shows the purchase rate as the middle number of each node. This starts out at 47% for all patients. The bottom number is the percentage of customers. The leaf on the far-left side accounts for 52% of all customers, a high percentage. This occurs from May – August or in November and December when interest rates rise above 2.5%. These patients have a low purchase rate of 25%.

In contrast, the leaf on the far-right side accounts for 25% of all patients. These people have a purchase rate that is much higher (85%). This occurs in April or May when the interest rate falls below 2.5%.

When ABC Insurance is marketing their products, they should prioritize customer contact during April and May when interest rates are below 2.5% and they should avoid calling customers during the months of May – August or November or December when interest rates are above 2.5%.

Task 12 – Choose a model (4 points)

Model	Test AUC
GLM with Age and Age^2 from Task 7	0.774
Elastic net with the same variables but smaller coefficients	0.7729
Decision tree from Task 10	0.7634
Pruned decision tree from Task 11	0.7613

I recommend using the pruned decision tree from Task 11. This has a reasonably high AUC and, more importantly, is straightforward to interpret. This GLM has the best AUC but includes many variables, including a principal component. The elastic net reflects an identical problem with its large number of variables. The unpruned decision tree has 14 nodes, too many for a marketing department to easily understand. Therefore, the pruned tree strikes a reasonable balance between predictive accuracy and ease of interpretation.

Task 13 – Executive summary (20 points)

The marketing department at ABC Insurance has requested our assistance with their phone marketing campaigns. We are helping them to increase their profits by improving their conversion rates, the percentage of customers who purchase a product after being contacted. In addition, we are interested in increasing the amount of insurance coverage purchased.

For each of the 10,000 phone calls, we have information regarding customer demographics, including age, years of education, marital status, job, housing, loan history, the day/month of the call, the type of phone, and the state of the economy. We are also informed whether the customer made a purchase.

The following questions can be used to determine or not a customer will make a purchase:

- **Is the interest rate greater than 2.5%?**
    - During May – August and November – December, the purchase rate is **lowest** at **25%**, reflecting 52% of all customers. We may conclude from this data that that ABC should not spend resources calling customers when the interest rates are below 2.5% during these months.
    - There were a few customers whose records appeared in other months, but they had a high purchase rate.
- **Is the interest rate less than 2.5%?**
    - During April and May, the purchase rate is **highest** (85%). ABC should prioritize calling customers at these times.
    - During April or May, if the CPI is more than 93, the purchase rate is **high** (66%).
    - During April or May, if the CPI is less than 93 and the interest rate is less than 1.3, then the purchase rate is **high** (72%).
    - At other times, the purchase rate is **low** (40%) and therefore we do not advise calling customers.

We began with an inspection of the quality of the data and adjusted it as needed. We wanted to ensure that this analysis would be based on real customer behavior and would be unaffected by data errors. All variables used matched the descriptions provided.

A few observations based on graphs and summary statistics include:

- The number of calls is uneven and depends on the month made.
- Data shows that most customers have a four-year college education.
- There is a close relationship between age and the customers' jobs. Students (median age of 25) tend to be younger while retired people (median age of 65) are older. Older customers are likely to be more comfortable speaking on the phone than through email or text, although a follow up analysis is necessary to confirm this hypothesis.
- Most often, insurance was purchased by younger people (under the age of 25) or older people (over the age of 40) . We recommend marketing to people immediately after they graduate from college and are no longer covered by their parent's insurance coverage, as well as older people who are thinking ahead to retirement.

It was important to consider educational level. This includes not only the type of education, but the total number of years including middle-high school, college, and post-graduate studies. For each person, we calculated total years of education and used this data in the analysis.

There were some missing values that needed to be fixed. In some cases, it was acceptable to simply remove the records with missing values, but in other cases this would have introduced survivorship bias into the analysis. As defined in this context, survivorship bias fails to consider what happened to customers not included in the study, basing conclusions solely on those with complete records. For example, customers given bad service might have hung up the phone before the call was concluded and thus never have their information recorded. These would be a lost sale for ABC Insurance, and so we do not want to ignore this information. To address this issue, we only removed values when we deemed that they were missing because of random chance, impacting a small number of customers. As an alternative, we filled in numeric values using averages and replaced categorical records using a special category.

We helped ABC Insurance to simplify their data by conducting a correlation analysis. Two things are correlated when increases/decreases in one tend to result in similar patterns in another. This data contained several measures of the state of the economy such as the consumer price index (CPI), the consumer confidence index (CCI), and short-term interest rates (irate). We confirmed that there is a correlation between irate and the number of employees of ABC Insurance, and irate and CPI. The body of this report contains the calculated correlations and may be reviewed if additional details are desired.

There is danger in using variables which are correlated directly in models. One way to address this problem is with a principal component decomposition. This merges variables which are correlated into fewer variables which contain most of the same information. We used this method to combine the economic variables into a single index, a numeric measure of the economy, which ABC can use in their analysis. We tested our models using both this new index as well as the original variables. One advantage to using the index is that the models may be more reliable. A disadvantage is that they are more difficult to understand.

To ensure that the results of this study can be reproduced in real life, we split the data into two groups: a training group where the models were built, and a testing group where the models were evaluated. There was no overlap between these groups of customers, and we could therefore simulate how reliable the results would be if used on new data.

We looked different types of models and considered their advantages and disadvantages.

- GLMs (Generalized linear models)
    - This model had the best performance but was also the most difficult to understand.
    - We found that a person's age is predictive of whether they will purchase or not.
- GLMs with stepwise selection
    - We tried making the above model simpler by using method known as stepwise selection. This made the model simpler and could be an alternative to the decision tree. One of the advantages of this model is that it can be explained in a spreadsheet. One disadvantage is that it used the results of the principal component analysis and so was complex to understand.
- GLMs with shrinkage

    - One reason for using this model is that it can simplify results by removing variables. However, in this case we did not find that any variables had been removed.
- Decision trees
    - The first trees that we fit were too complex to understand.
    - Trees were too sensitive to the data used to build them and did not identify patterns which could be generalized to new data.
- Pruned decision trees
    - This model struck a balance between good performance while also being explainable.
    - Handled low-quality and complex data automatically which can be useful for future studies. In this case, we needed to replace missing values by hand and identify complex data relations by looking at graphs, a time-consuming process.
    - By using a tree-based model, we could perform these steps quickly when analyzing different customers.

In conclusion, this report shows that using predictive analytics might offer important clues to whether a customer will purchase, even in advance of a phone call. Marketing managers can examine customer demographic data as well as the environment and use this information to optimize conversion rates. The results of this study were based on limited data from the year 2014. Therefore, if the company wants to apply these conclusions to future years, new data will need to be supplied and this analysis updated. We assumed that records which were removed had randomly missing data. However, a pattern existing behind these missing values could materially change the findings. In addition, details of actual phone conversations, such the way in which questions are asked, can strongly affect the answers. We have assumed that ABC's phone call tactics will not change. Any changes in these procedures may also affect the results cited here.