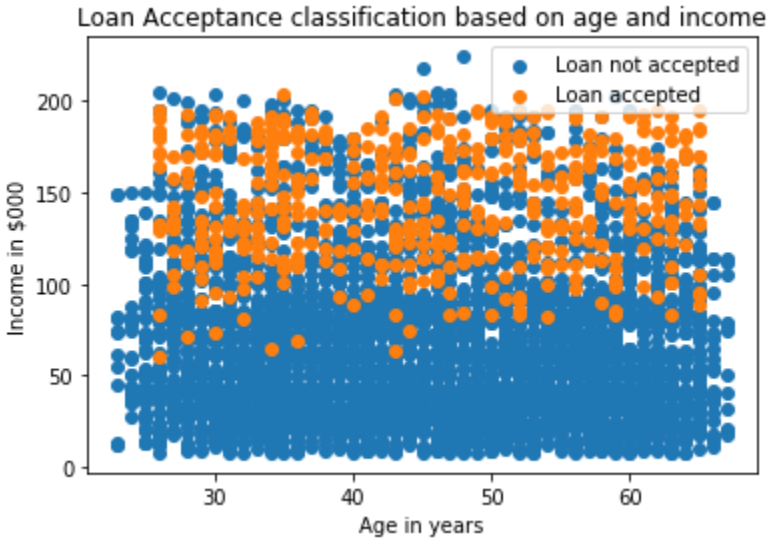
Homework 2

1. **Create a scatterplot of Age vs. Income, using color to differentiate customers who accepted the loan and those who did not. Which variable (i.e., age or income) appears to be potentially more useful in classifying customers? Explain.**



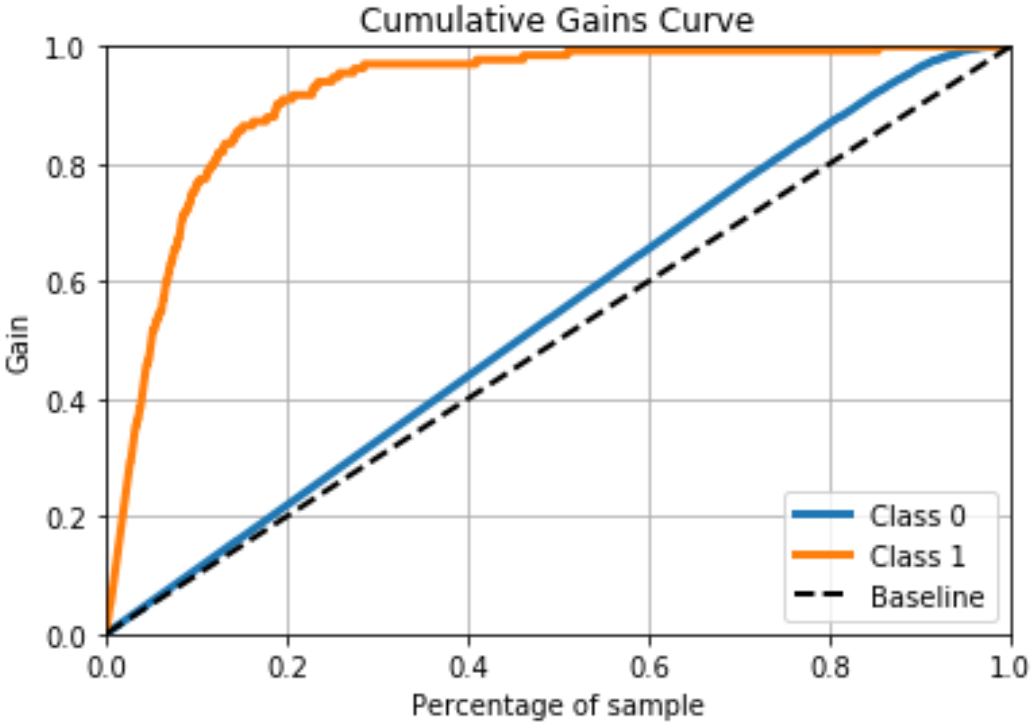
From the above plot, it is very clear that Income variable appears to be potentially more useful in classifying customers. Apart from a few exceptions that have the lowest income below 1,00,000$, rest all the customers who have accepted the personal loan lie above the $1,00,000 mark. Similarly, the region below $100000 income have most of the customers who have not accepted the personal loan.

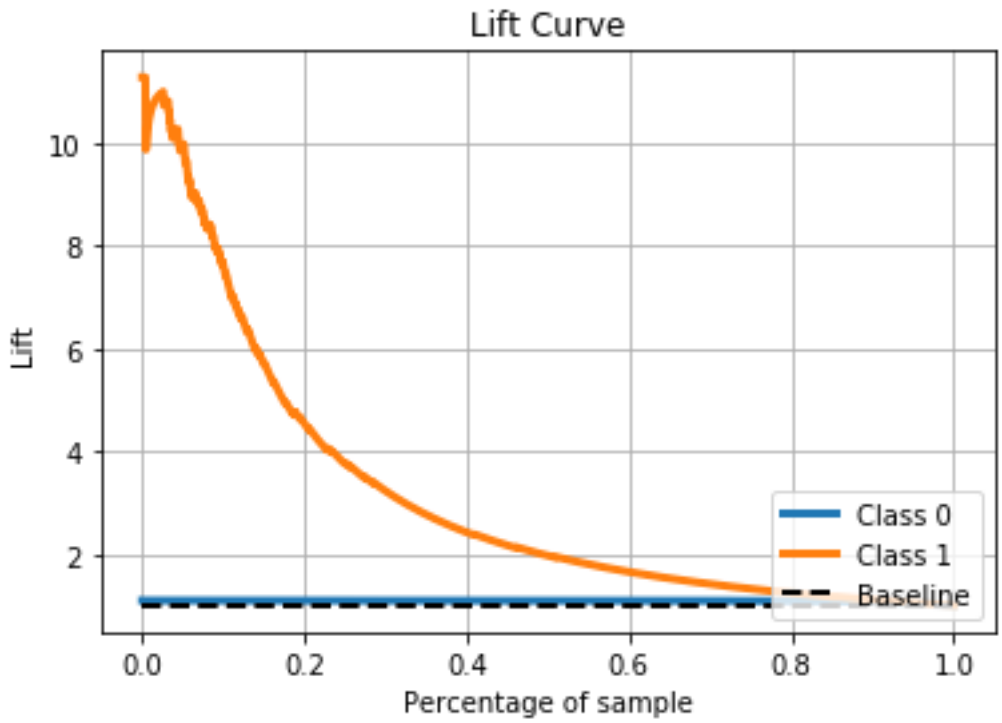
2. **Build a logistic regression model to classify customers into those who are likely to accept personal loan offer and those who are not. Use all the available variables as predictors except ID and ZIP Code. (Hint: Since the Logistic Regression operator expects binominal or polynominal target variables, if the target variable is numeric, you will have to convert it to binominal by using the Numerical to Binominal operator.)**
   a. **Evaluate the overall predictive accuracy of the model as well as the accuracy of each class using appropriate metrics.**
      The overall predictive accuracy of the model is 94.33%
      The accuracy of positive class (i.e., customers who accepted the loan offer) is 54.88%
      The accuracy of negative class (i.e., customers who did not accepted the loan offer) is 98.82%
   b. **What was the default cutoff probability used to generate the classifications?**
      The default cutoff probability used to generate the classifications is 0.50 i.e., predicted probability above 0.5 was classified as class label 1 and predicted probability below 0.5 was classified as class label 0.
   c. **Assuming that the dataset contains a representative sample of the liability customers of the bank, if you target 100 customers randomly (i.e., without the**

aid of any predictive model), how many of them would potentially accept a personal loan offer?

If we take 100 customers randomly from the dataset, then approximately 10 customers accept the personal loan offer.

d. **Now if you use your model in Part (2) to select 100 customers with the highest probability of loan acceptance, how many of them would potentially accept a personal loan offer? (Hint: Revise the process from Part (2) to generate a lift chart.)**





The lift chart shows how much more likely we are to receive positive responses than if we contact a random sample of customers. For example, by contacting only top 10% of customers (100 customers) based on the predictive model we will reach 10 times as many respondents, as opposed to no model. If we use random sampling, we get approx. 10 customers who accept the loan as compared to first 100 customers with highest probability who accept almost 10 times as that of baseline model.

3. **Suppose the bank is interested in improving the accuracy of identifying the potential positive responders, i.e., those who would accept the loan offer. Create a new process to develop a logistic regression model to classify customers into those who are likely to accept personal loan and those who are not using all the available variables—except ID and ZIP Code — as predictors. However, this time modify the cutoff probability in such a way that the accuracy of identifying the positive responders is at least 70%. Compare the predictive accuracy of this revised model with that of the model developed in Part (2). (Again, try to be analytical instead of just noting the numbers)**

When we adjusted the test cutoff from 0.5 to 0.3, we have increased our models accuracy by 19% up from 54% to 73%. The number of positive responders is less compared to those who do not accept the personal loan offer. Due to this, if we initially keep the default probability to 0.5 we get accuracy of positive responders as 54% because the number of positive responders below is threshold value is less. Hence, we lower the cutoff probability to 0.3 from 0.5, we will be able to get more number of positive responders that satisfy the condition:- If probability of predictive variable is greater than

0.3 then label the class as 1 (positive responder) else label the class as 0. This increases the percentage of positive responders as 73%.

4. **Aside from the problem of predicting the likelihood of accepting loan offers, think of two other business problems where logistic regressions can be utilized for predictive modeling. For each problem, identify a target variable and four possible predictor variables.**

   Two examples of business problems where logistic regression could be used are –

   1. Healthcare facility system can use logistic regression to find if a diagnosis of patients who have a tumor is malignant or benign. The target variable will be a diagnosis variable, where output can be either malignant or benign. The four possible predictor variables would be radius, texture, perimeter, area of the tumor.
   2. Gmail Spam detection -
      The target variable would be whether the email is spam or ham. The four possible predictor variables would be text from email, subject of email text, mail client, subject contains special characters.