Name: Sri Vamshi Polela

CSDA 5430 Predictive Analytics

Week 3 Assignment

1. A manufacturer of an electromechanical kitchen utensil conducted an analysis of a large number of consumer complaints and found that they fell into the six categories shown in the following table. Based on numbers in the table, calculations of following probabilities.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Distribution of Product Complaints** | | | | |
|  | Reason for Complaint | | | |
| Electrical % | Mechanical % | Appearance % | Totals % |
| During Guarantee Period | 18 | 13 | 32 | 63 |
| After Guarantee Period | 12 | 22 | 3 | 37 |
| Total | 30 | 35 | 35 | 100 |

* What is the probability of electrical related complaints? What is the probability of complaints that were after guarantee period?
* Probability of Electric related complaints**:**

= Total number of Electrical complaints / Total number of complaints

= 30 / 100 = 0.30

* Probability of complaints that were after guarantee period:

Total number of After Guarantee Period / Total number of complaints

= 37 / 100 = 0.37

**(Above are good)**

* What is the probability of complaints that were mechanical related and during guarantee period?

= (35 / 100) \* (13 / 35)

= 0.35 \* 0.37

= 0.1295

**(Calculation error. The joint probability showing on the distribution is 0.13.)**

* If a consumer complaint is received, what is the probability of mechanical related complaints given that the complaints are for during guarantee period?

= (13 / 35 \* 35 / 100) / (63 / 100)

= (0.37 \* 0.35) / 0.63

= 0.2055 = 0.26

**(0.2055 is not 0.26. P(mechanical-related | during-guarantee) = 13/63 = 0.206.)**

* If a consumer complaint is received, what is the probability that the complaint was originated after the guarantee period given that the cause of the complaint was not product appearance?

= P(Not Product Appearance | After guarantee period) \* P(After guarantee period) / P(Not Product Appearance)

= ((22 + 12) / 37 \* 37 / 100) / ((30 + 35) / 100)

= ((34 / 37) \* 0.37) / (65 / 100) = 0.34 / 0.65

= 0.52 **(This is okay, but unnecessary complex. All calculation can be directly come from the distribution table.)**

**8/10 points**

1. What are the practical difficulties with complete (exact) Bayes algorithm in calculating conditional probabilities? Why is the naïve assumption, assuming independent events, necessary for naïve Bayes algorithm? Use examples in chapter 7 case study (head and neck cancer medication) to support your discussion.

| **Difficulty** | **(Exact) Bayes Algorithm** | **Naive Bayes Algorithm** |
| --- | --- | --- |
| Computational Complexity | Becomes computationally complex with many variables and data points | Reduces computational complexity by assuming independence between variables |
| Data Sparsity | Requires a large amount of data to estimate joint probability | Can handle sparse data more effectively due to simplified computation |
| Model Specification | Requires specifying a full probabilistic model | Assumes independence between variables given the class label |
| Storage Requirements | Memory-intensive | Requires less memory |

* **In the context of the head and neck cancer medication case study:**
* **(Exact) Bayes Algorithm:** Calculating conditional probabilities for each combination of variables related to medication, dosage, frequency, and cancer stage would be impractical due to computational complexity and data sparsity.
* **Naïve Bayes Algorithm:** Assumes independence between medication variables, dosage, frequency, and cancer stage, allowing for efficient computation of conditional probabilities despite the complexities of the dataset. **(Not well discussed. Frequency and dosage are not predators for the case. The variables or predictors for this case study are various words/terms that are extracted from text mining process. With the naïve assumption, various predictors (in this case, various words/terms) are independent, meaning a word occurrence is not affect another. However, words do go together often, like “every day” and “every x hours”. With this simplified assumption of conditional independence, exact conditional probability is approximated by the product of all the individual conditional probabilities – probabilities of each word/term.)**

**6/10 points**

1. In data analytics, rescaling can make each feature contribute to the analysis in a relatively equal manner, avoiding potential bias. In R, perform the following activities on below customer data.

|  |  |  |  |
| --- | --- | --- | --- |
| Record | Age | Income ($) | Female |
| 1 | 25 | 49,000 | 0 |
| 2 | 56 | 156,000 | 1 |
| 3 | 65 | 99,000 | 1 |
| 4 | 32 | 192,000 | 0 |
| 5 | 41 | 39,000 | 0 |
| 6 | 49 | 57,000 | 0 |

* Rescale the data (ignore record number) in both z-score standardization and min-max normalization.
* **After rescaling the results are below:**

A screenshot of a computer

Description automatically generated

**(Looks good)**

* **For before and after rescaling, compute distance matrix (may use *dist()* function) and identify pairs with minimum and maximum distances.**

A screenshot of a computer

Description automatically generated A screenshot of a computer code

Description automatically generated

**(It is incorrect to remove the first column when calculate distance for original data. Others are good.)**

* **Comment on the impact of rescaling.**
* Z-Score Standardization: It scales the data by subtracting the mean and dividing it by the standard deviation.
* Min-Max Scaling: It scales the data by transforming it to a specific range, usually [0, 1].

**Impact:**

* Z-Score Standardization method is beneficial for algorithms relying on distance metrics and maintaining the original distribution's shape.
* Min-Max Scaling preserves relative differences and reduces the impact of outliers, it may not be suitable for algorithms assuming normal distribution or relying on specific statistical assumptions. **(What do you see before and after rescaling on the data? What is the impact of rescaling? The impact can be observed for before and after rescaling, the pairs with min and max distances are changed, which means distance distribution among records is different.)**

**7/10 points**

1. Data file *OnlineShopper.csv* contains browsing behavior on various websites for 29 online shoppers. A naïve Bayes model was built on the data as shown below.

A screenshot of a computer

Description automatically generated

* Explain each number under a-priori probabilities and conditional probabilities for *sports.com*.

1. **priori probabilities**

Probability of the target variable being "No" is approximately 55.17%.

Probability of the target variable being "Yes" is approximately 44.83%.

**(Good)**

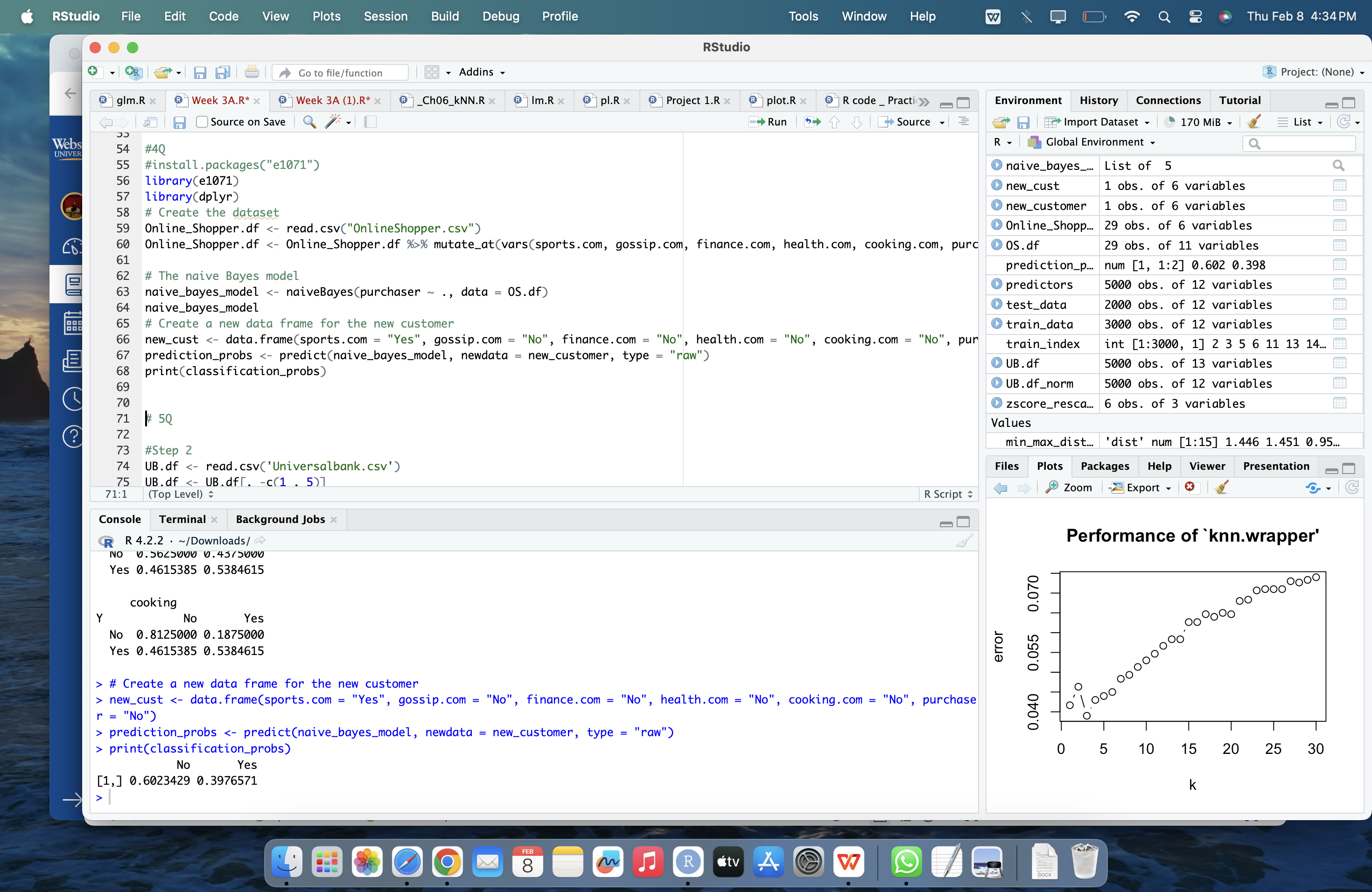
**Conditional probabilities:**

Given that the target variable is "No", the probability of having "No" in the variable "sports.com" is approximately 81.25%, and the probability of having "Yes" is approximately 18.75%.

Given that the target variable is "Yes", the probability of having "No" in the variable "sports.com" is approximately 30.77%, and the probability of having "Yes" is approximately 69.23%.

**(good)**

* Discuss naïve assumptions in this model.
* In a broader context of naive Bayes models, the assumption of independence between predictor variables given the target variable is typically made, simplifying the modeling process.
* Despite this simplification, naive Bayes models often prove effective due to their computational efficiency and practical utility in handling large datasets. **(What does it mean in this model?** **Each predictor represents an event happen or not happen. For example, sports.com represents a customer visit the website or not. The naïve assumption is that customers visiting a website is independent to visiting other websites. Per the assumption, various websites visited by shoppers are independent and have nothing to do with each other. This is not true in reality. For example, shoppers having interest in health.com may be more likely to visit cooking.com because cooking and health are related. Shoppers visited sports.com may be more interested in knowing stories behind the scenes so they are more likely to visit gossip.com.)**
* Make a classification on the new customer who visited *sports.com* website but did not visit any others. Explain how it is classified.

**

* Seeing the output, it can be classified that the new customer would be classified to non-purchaser

**(The customer would be classified as non-purchaser because naïve probability for non-purchase is higher than that for purchaser.)**

**8/10 points**

1. Universal Bank is a relatively young bank growing rapidly in terms of overall customer acquisition. The majority of these customers are liability customers (depositors) with varying sizes of relationship with the bank. The customer base of asset customers (borrowers) is quite small, and the bank is interested in expanding this base rapidly to bring in more loan business. In particular, it wants to explore ways of converting its liability customers to personal loan customers (while retaining them as depositors).

A campaign that the bank ran last year for liability customers showed a healthy conversion rate of over 9% success. This has encouraged the retail marketing department to devise smarter campaigns with better target marketing. The goal is to predict whether a new customer will accept a loan offer. This will serve as the basis for the design of a new campaign.

The file *UniversalBank.csv* contains data on 5000 customers. The data include customer demographic information (age, income, etc.), the customer’s relationship with the bank (mortgage, securities account, etc.), and the customer response to the last personal loan campaign (Personal Loan). Among these 5000 customers, only 480 (= 9.6%) accepted the personal loan that was offered to them in the earlier campaign.

In R, perform the following steps as discussed in Ch06 on the data.

* Step 1: Collecting data (Discuss the business problem and how the data can support the analytics, identify the important class.)

**Business problem:**

* Universal Bank has 5000 customers but most of them are depositors. The bank decided to bring some ways to convert them as personal loan customers (while retaining them as depositors).
* Banks want to increase the personal loan business and the goal is to predict whether a new customer will accept a loan offer or not.

**Data Understanding:**

* **Key variables** include age, income, mortgage, securities account,
* The Important class for analysis is customers who accepted personal loans (Personal.Loan=1). Focusing on this class is pivotal in designing effective campaigns targeted at converting liability customers into personal loan customers. The primary goal is predicting whether a new customer will accept a loan offer, making accurate identification of loan acceptors more crucial than non-acceptors. **(Okay)**
* **Step 2:** Exploring data, rescaling data, creating training and test datasets (Remove zipcode, create dummies for education. Rescaling the data. Create training partition (60%) and testing partition (40%) with randomized partitioning. Set a seed so the results can be reproduced. What is the percentage of customers who accepted the personal loan? How about this percentage in training and testing partitions, respectively?)

A screenshot of a computer

Description automatically generated

**(The training data is invalid. First, it has education as well as education dummies. It also contains personal loan.)**

* Step 3: Training a model on the data (Use the function knn() from the class package for the kNN algorithm. Try with any initial number for k. Remember use all dummy variables for knn and ignore customer id as part of predictors.)

A black text on a white background

Description automatically generated

**(Invalid data and invalid model. Why is the last variable removed by such as in training data -ncol(train\_data)]?)**

* Step 4: Evaluating model performance (Evaluate the model performance with confusion matrix. Identify accuracy, sensitivity, and specificity.)

A screenshot of a computer

Description automatically generated

**(Invalid data and invalid model.)**

* Step 5: Improving model performance, testing alternative values of k (Use tune.knn() function to identify the best k, test with the best k and compare the model performance with the initial k. Optionally, visualize the error rate with training, CV, and testing error rate against k.)
* K =1 is best

**A white screen with black text

Description automatically generated**

* **After tunning the model we got k=1 and again running the model with k=1 sensitivity is increased by 10%**

**A screenshot of a computer

Description automatically generated**

**(Invalid data for model.)**

**10/20 points**

**Total 39/60 point**

**Description of Variables in Universal Bank Dataset**

|  |  |
| --- | --- |
| **Variable** | **Description** |
| ID | Customer ID |
| Age | Customer's age in completed years |
| Experience | #years of professional experience |
| Income | Annual income of the customer ($000) |
| ZIPCode | Home Address ZIP code. |
| Family | Family size of the customer |
| CCAvg | Avg. spending on credit cards per month ($000) |
| Education | Education Level. 1: Undergrad; 2: Graduate; 3: Advanced/Professional |
| Mortgage | Value of house mortgage if any. ($000) |
| Personal Loan | Did this customer accept the personal loan offered in the last campaign? |
| Securities Account | Does the customer have a securities account with the bank? |
| CD Account | Does the customer have a certificate of deposit (CD) account with the bank? |
| Online | Does the customer use internet banking facilities? |
| CreditCard | Does the customer use a credit card issued by UniversalBank? |