# Naive Bayes

**Business problem:** Build a naive Bayes model on the data set for classifying the ham and spam.

**Data:**

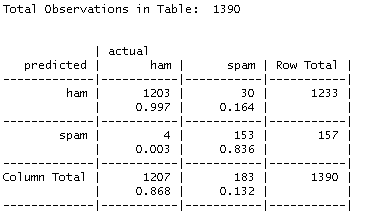
Clean the dataset and tokenize them.

* Converted them into lower cases.
* Removed the punctuations marks.
* Removed numbers.
* Removed stop words.

**Predicting the Model:**

|  |  |  |
| --- | --- | --- |
| Ham | 1233 | .0887 |
| Spam | 157 | .0113 |

**Output of the Model:**

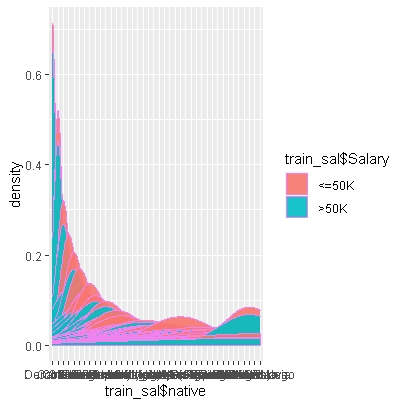
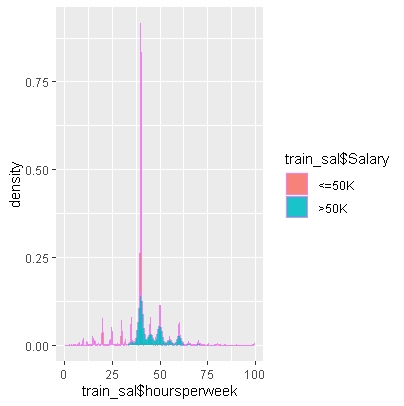
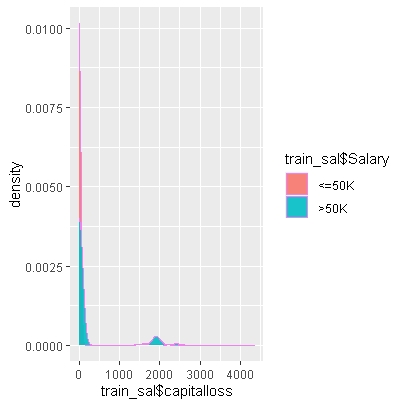
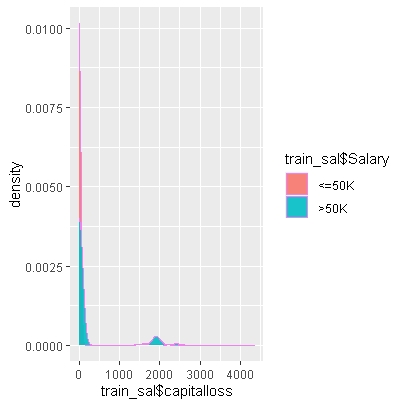
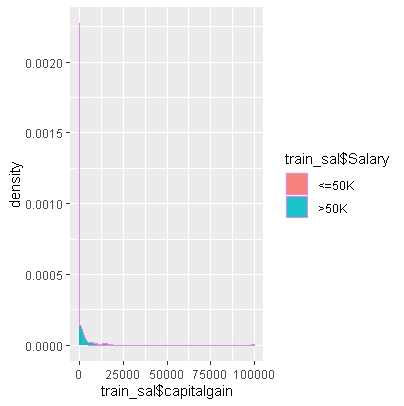
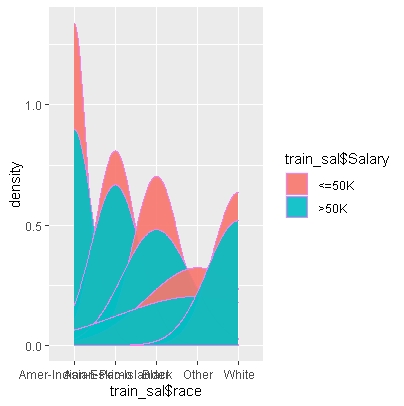
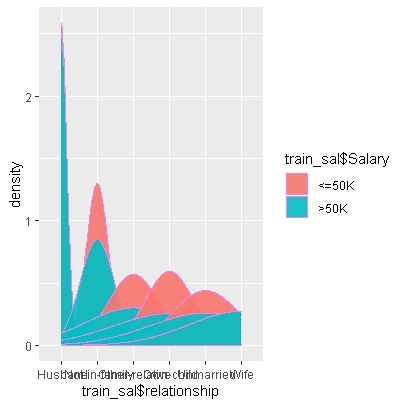
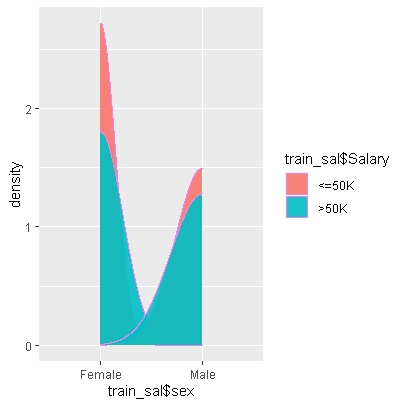
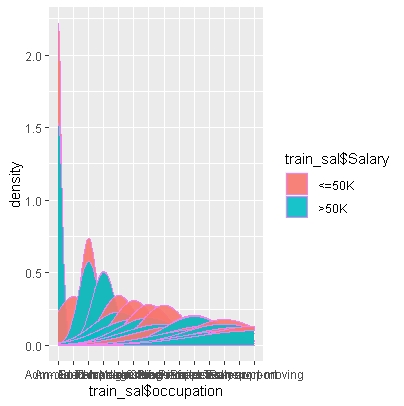
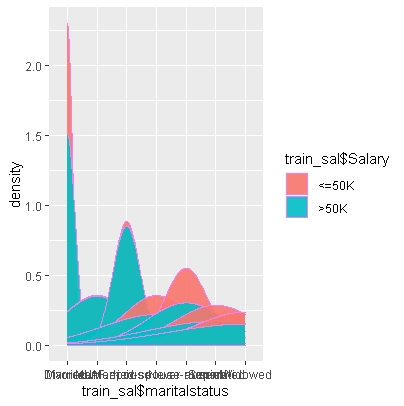
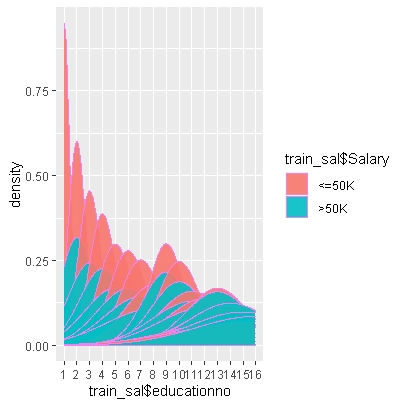
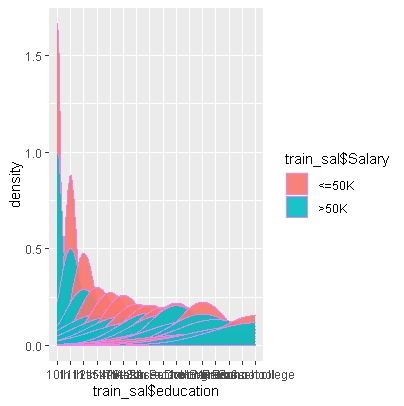
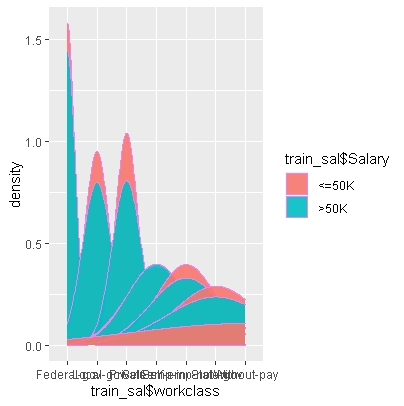
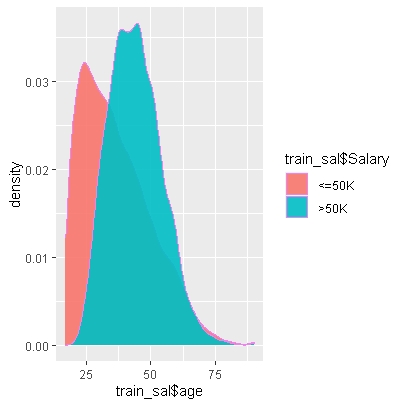
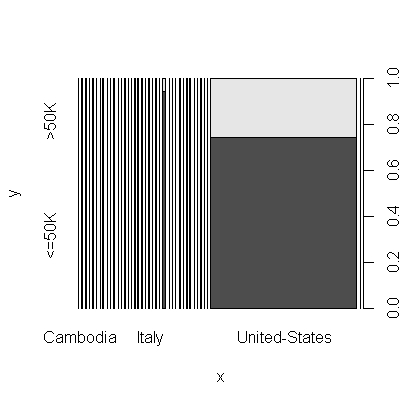
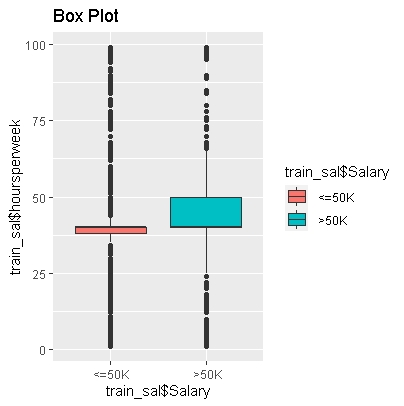
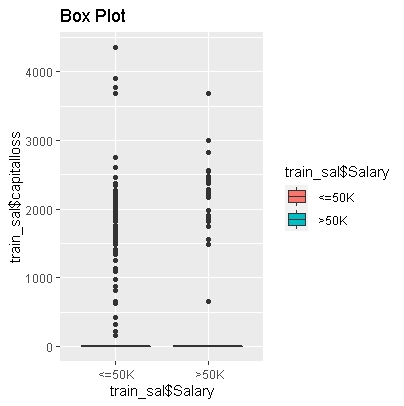
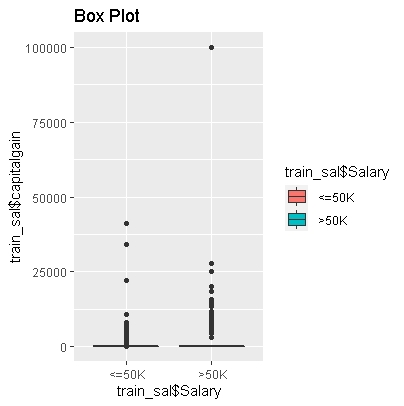
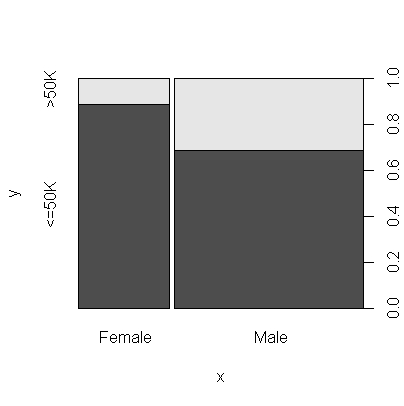
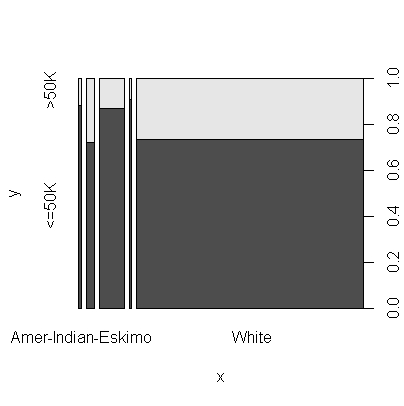
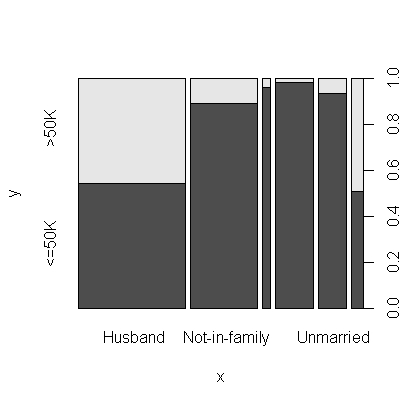
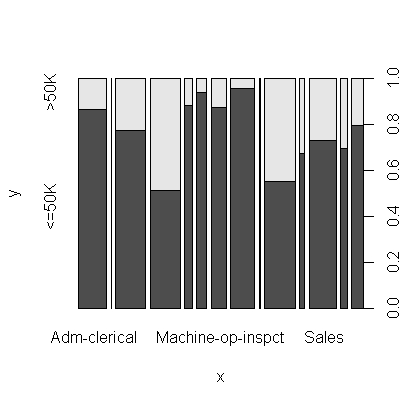
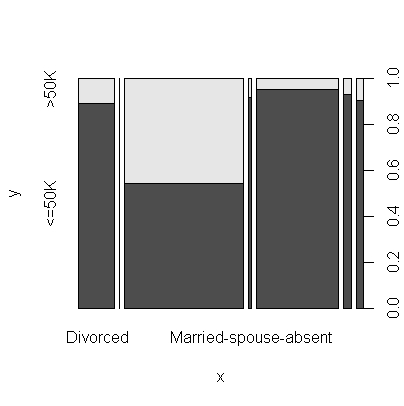
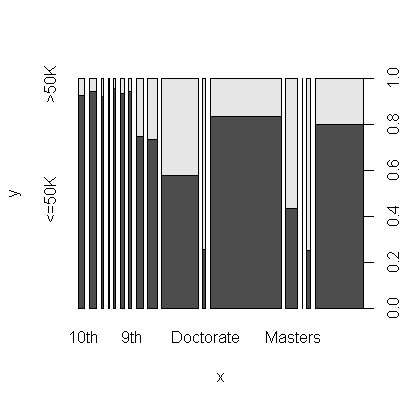
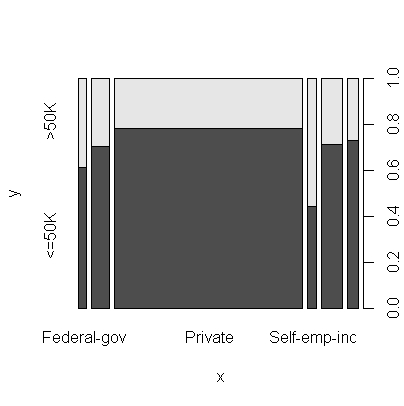
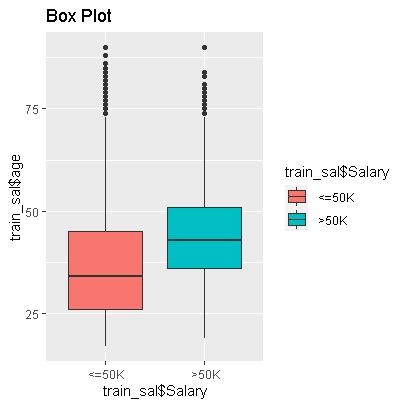


**Business** **Problem**: Prepare a classification model using Naive Bayes for salary data.

**Data:** The dataset has 45221 observations and has 14 features.

**Data Visualization:**

Attached is the different type of plotting for all columns reference with Salary feature.

****

Apply Naive Bayes Classifier for Discrete Predictors:

Call:

naiveBayes.default(x = X, y = Y, laplace = laplace)

A-priori probabilities:

Y

<=50K >50K

0.7510693 0.2489307

Conditional probabilities:

age

Y [,1] [,2]

<=50K 36.60826 13.46489

>50K 43.95911 10.26963

workclass

Y Federal-gov Local-gov Private Self-emp-inc Self-emp-not-inc State-gov Without-pay

<=50K 0.0255153843 0.0643623361 0.7685074825 0.0209243809 0.0787975103 0.0412748863 0.0006180197

>50K 0.0486148109 0.0811134790 0.6494405967 0.0799147576 0.0950985615 0.0458177944 0.0000000000

education

Y 10th 11th 12th 1st-4th 5th-6th 7th-8th 9th Assoc-acdm Assoc-voc

<=50K 0.0335937845 0.0436586766 0.0153622037 0.0064009182 0.0121838167 0.0230433055 0.0189820333 0.0331964861 0.0425109257

>50K 0.0078582845 0.0078582845 0.0038625466 0.0007991476 0.0015982952 0.0046616942 0.0033297816 0.0340969632 0.0458177944

education

Y Bachelors Doctorate HS-grad Masters Preschool Prof-school Some-college

<=50K 0.1288129608 0.0041937050 0.3629982784 0.0312982828 0.0019864919 0.0060036198 0.2357745111

>50K 0.2831646244 0.0372935535 0.2153702717 0.1222695791 0.0000000000 0.0540756526 0.1779435269

educationno

Y 1 2 3 4 5 6 7 8 9

<=50K 0.0019864919 0.0064009182 0.0121838167 0.0230433055 0.0189820333 0.0335937845 0.0436586766 0.0153622037 0.3629982784

>50K 0.0000000000 0.0007991476 0.0015982952 0.0046616942 0.0033297816 0.0078582845 0.0078582845 0.0038625466 0.2153702717

educationno

Y 10 11 12 13 14 15 16

<=50K 0.2357745111 0.0425109257 0.0331964861 0.1288129608 0.0312982828 0.0060036198 0.0041937050

>50K 0.1779435269 0.0458177944 0.0340969632 0.2831646244 0.1222695791 0.0540756526 0.0372935535

maritalstatus

Y Divorced Married-AF-spouse Married-civ-spouse Married-spouse-absent Never-married Separated Widowed

<=50K 0.1660707191 0.0004855869 0.3384099236 0.0149649053 0.4085551583 0.0385379420 0.0329757648

>50K 0.0602024507 0.0013319126 0.8522908897 0.0041289291 0.0625998934 0.0087906233 0.0106553010

occupation

Y Adm-clerical Armed-Forces Craft-repair Exec-managerial Farming-fishing Handlers-cleaners Machine-op-inspct

<=50K 0.1422769611 0.0003531541 0.1378183905 0.0907164614 0.0385820863 0.0559307818 0.0759281331

>50K 0.0663292488 0.0001331913 0.1209376665 0.2579914758 0.0153169952 0.0110548748 0.0326318594

occupation

Y Other-service Priv-house-serv Prof-specialty Protective-serv Sales Tech-support Transport-moving

<=50K 0.1359643314 0.0062684854 0.0983092747 0.0191586103 0.1153931047 0.0279874630 0.0553127621

>50K 0.0175812467 0.0001331913 0.2412093767 0.0279701652 0.1291955248 0.0370271710 0.0424880128

relationship

Y Husband Not-in-family Other-relative Own-child Unmarried Wife

<=50K 0.299474683 0.304727851 0.037655057 0.194323048 0.132388646 0.031430716

>50K 0.756393181 0.109616409 0.004661694 0.008524241 0.028369739 0.092434736

race

Y Amer-Indian-Eskimo Asian-Pac-Islander Black Other White

<=50K 0.011124354 0.028561338 0.108197590 0.009270295 0.842846422

>50K 0.004528503 0.033031433 0.048748002 0.002797017 0.910895045

sex

Y Female Male

<=50K 0.3826866 0.6173134

>50K 0.1481087 0.8518913

capitalgain

Y [,1] [,2]

<=50K 148.9004 936.4124

>50K 3937.6798 14386.0600

capitalloss

Y [,1] [,2]

<=50K 53.35302 309.9476

>50K 193.75067 592.8256

hoursperweek

Y [,1] [,2]

<=50K 39.34856 11.95104

>50K 45.70658 10.73699

native

Y Cambodia Canada China Columbia Cuba Dominican-Republic Ecuador El-Salvador

<=50K 0.0004855869 0.0031342427 0.0021189246 0.0023837902 0.0029576657 0.0028693771 0.0010153181 0.0040171280

>50K 0.0009323388 0.0047948855 0.0026638253 0.0002663825 0.0033297816 0.0002663825 0.0005327651 0.0011987214

native

Y England France Germany Greece Guatemala Haiti Honduras Hong Hungary

<=50K 0.0024720788 0.0006621640 0.0037081181 0.0009270295 0.0026486558 0.0016774820 0.0004855869 0.0005738754 0.0004414426

>50K 0.0039957379 0.0015982952 0.0058604156 0.0010655301 0.0003995738 0.0005327651 0.0001331913 0.0007991476 0.0003995738

native

Y India Iran Ireland Italy Jamaica Japan Laos Mexico Nicaragua

<=50K 0.0026486558 0.0010594623 0.0008387410 0.0019423476 0.0030900984 0.0015891935 0.0006621640 0.0254712400 0.0013684722

>50K 0.0053276505 0.0023974427 0.0006659563 0.0031965903 0.0013319126 0.0030633990 0.0002663825 0.0043953117 0.0002663825

native

Y Outlying-US(Guam-USVI-etc) Peru Philippines Poland Portugal Puerto-Rico Scotland South

<=50K 0.0006180197 0.0012360394 0.0056504657 0.0019864919 0.0013243279 0.0042819936 0.0003972984 0.0025162230

>50K 0.0000000000 0.0002663825 0.0079914758 0.0014651039 0.0005327651 0.0015982952 0.0002663825 0.0018646777

native

Y Taiwan Thailand Trinadad&Tobago United-States Vietnam Yugoslavia

<=50K 0.0010153181 0.0006180197 0.0007063082 0.9053546992 0.0026045115 0.0004414426

>50K 0.0025306340 0.0003995738 0.0002663825 0.9316728823 0.0006659563 0.0007991476

**Confusion Matrix and Statistics :**

Reference

Prediction <=50K >50K

<=50K 10549 1919

>50K 811 1781

Accuracy : 0.8187

95% CI : (0.8125, 0.8248)

No Information Rate : 0.7543

P-Value [Acc > NIR] : < 2.2e-16

Kappa : 0.456

Mcnemar's Test P-Value : < 2.2e-16

Sensitivity : 0.9286

Specificity : 0.4814

Pos Pred Value : 0.8461

Neg Pred Value : 0.6871

Prevalence : 0.7543

Detection Rate : 0.7005

Detection Prevalence : 0.8279

Balanced Accuracy : 0.7050

'Positive' Class : <=50K