Statistical and Predictive Modeling II (DATA 2204)

Final Assignment

Professor: Fatma Tetikoglu

By : John Shaju

100852373

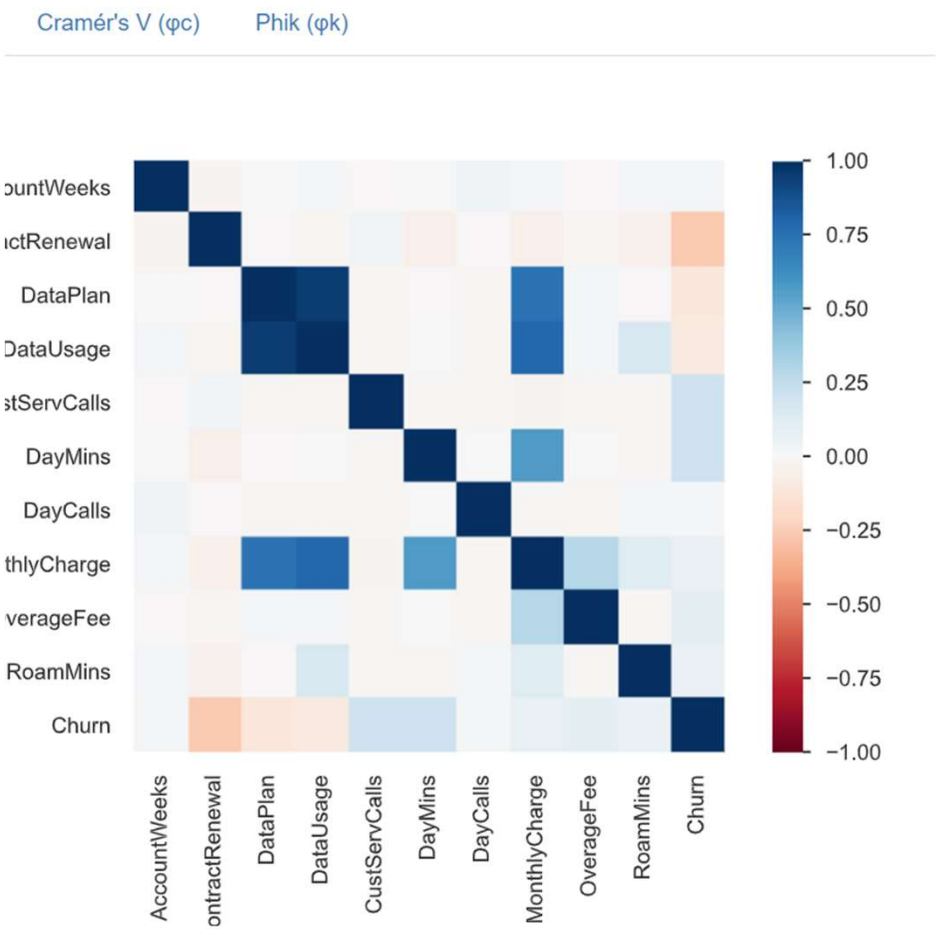
Problem Statement

- After evaluating the wireless churn.csv dataset, Mr. John Hughes has asked you to develop a forecasting model that incorporates both naive Bayes and logistic regression.
-

Dataset contains: 3,333 observations and 11 variables:
Independent Variables
AccountWeeks - number of weeks customer has had active account
ContractRenewal - 1 if customer recently renewed contract, 0 if not
DataPlan - 1 if customer has data plan, 0 if not
DataUsage - gigabytes of monthly data usage
CustServCalls - number of calls into customer service
DayMins - average daytime minutes per month
DayCalls - average number of daytime calls
MonthlyCharge - average monthly bill
OverageFee - largest overage fee in last 12 months
RoamMins – average roaming minutes per month
Dependent Variables
Churn - 1 if customer cancelled service, 0 if not

Exploratory Data Analysis

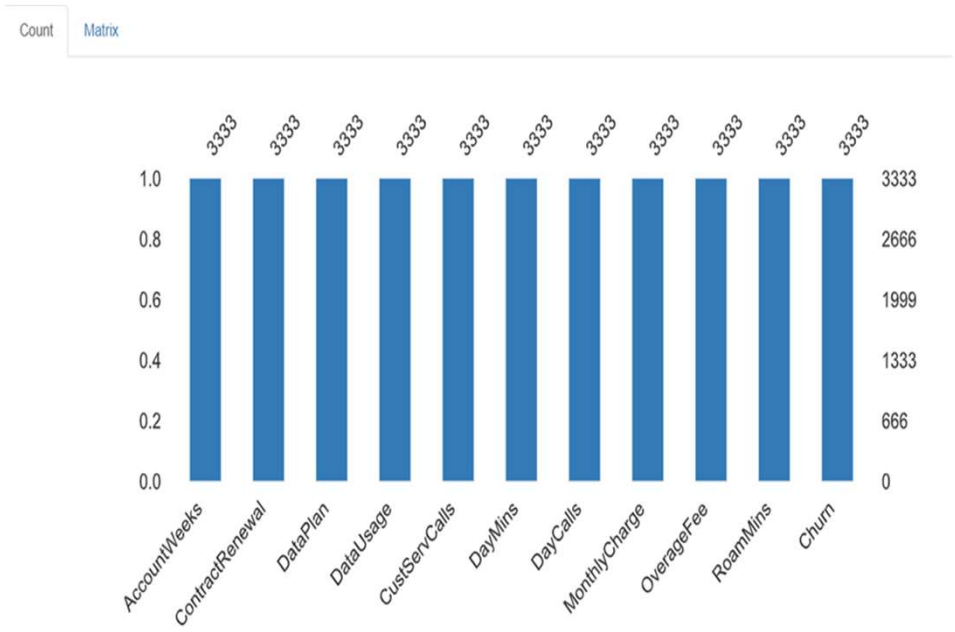
- Correlation : Data Usage and one additional field, as well as Data Plan, have a strong correlation. Data Usage and Data Plan and two additional fields have a strong correlation. Day Mins and Monthly Charge have a strong correlation. Monthly Charge and Data Plan have a strong correlation. Roam Mins and Data Usage have a strong correlation.



Exploratory Data Analysis

- Null Values : There are a total of 3333 observations and 11 variables in the dataset and there are no null values

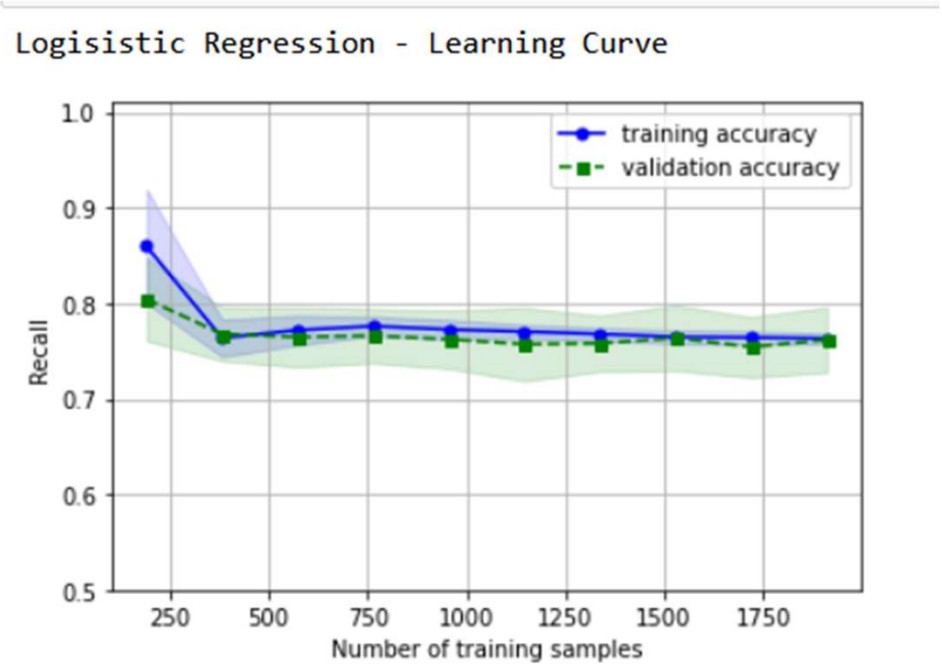
- Variables : there are 3 categorical variables and 8 numerical variables.



Learning Curve for Logistical Regression

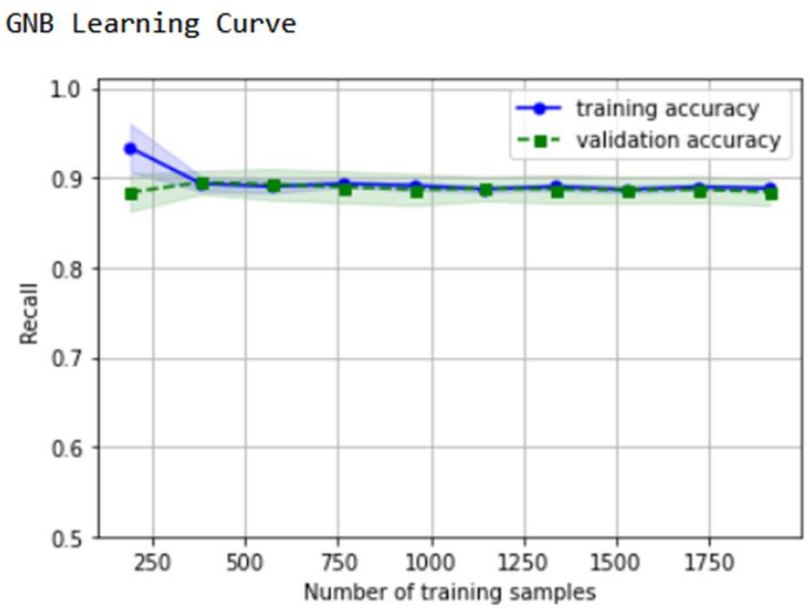
- The relationship between a person's level of competence and their experience is represented graphically by a learning curve. It is typical for proficiency measured on the vertical axis to rise with experience measured on the horizontal axis, i.e., the more often someone, groups, businesses, or sectors undertake an activity, the better they perform at it.
- The training recall curve starts from 80 and falls at 75 which is less than the average that indicate that the training curve is underfitting and it does not predict the data set properly .

- For the validation recall curve is lower than 75 which means there is less variability in the dataset which means there is less error in the test dataset



Learning Curve for Naïve Base

- The relationship between a person's level of competence and their experience is represented graphically by a learning curve. It is typical for proficiency measured on the vertical axis to rise with experience measured on the horizontal axis, i.e., the more often someone, groups, businesses, or sectors undertake an activity, the better they perform at it.
- The training recall curve starts from 89 and goes to 90 which is less than the average that indicate that the training curve is underfitting and it does not predict the data set properly .
- Here the training and validation curve are overlapping each other

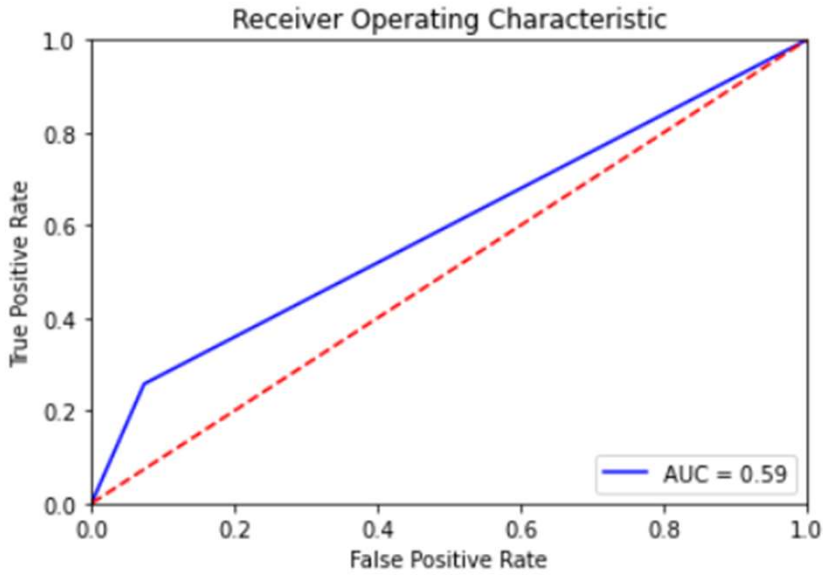


Classification report for Logistic Regression

- Precision : Being correct in your model is what precision is all about. In other words, you can quantify the likelihood that a model's predictions are accurate. In this instance, 88% accurate in predicting the no cheurn and 37% accurate in predicting with cheurn.
- Recall: The proportion of accurately anticipated positive observations to all the actual class's observations is known as recall. In this instance 93% of the model has classified as no cheurn outcome and 26% as accurate with cheurn outcome .
- F1 score : F1 score is the weighted harmonic mean of precision and recall. The macro average of the f1 score is 90% and the weighted average is 60%. The accuracy is 83%.

```
Model Name: LogisticRegression(class_weight='balanced', max_iter=1000, random_state=100)

Best Parameters: {'clf__C': 0.01, 'clf__penalty': 'l2'}

 [[528  42]
 [ 72  25]]

              precision    recall  f1-score   support

   Outcome 0       0.88      0.93      0.90       570
   Outcome 1       0.37      0.26      0.30        97

    accuracy                           0.83       667
   macro avg       0.63      0.59      0.60       667
weighted avg       0.81      0.83      0.82       667
```

ROC Curve

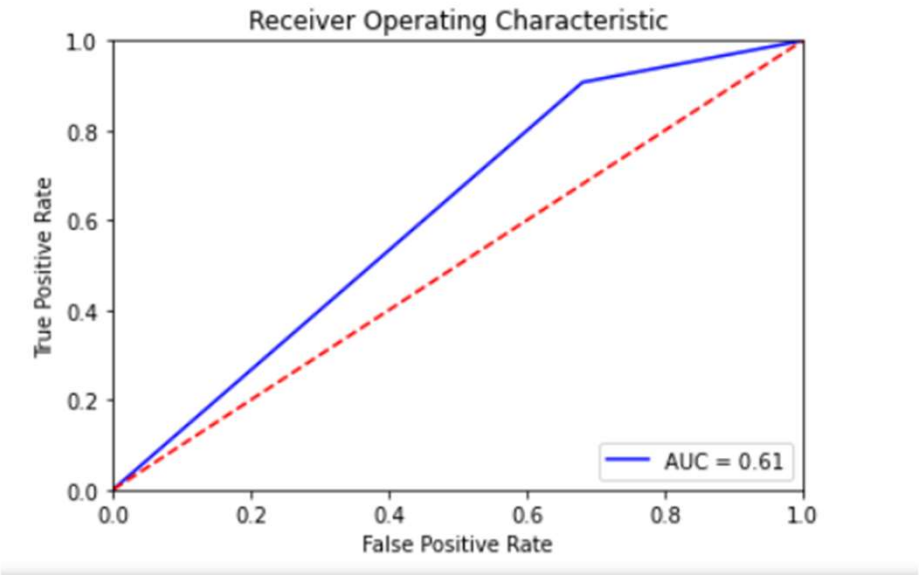


Classification report for Naïve Base

- Precision : Being correct in your model is what precision is all about. In other words, you can quantify the likelihood that a model's predictions are accurate. In this instance, 95% accurate in predicting the no cheurn and 18% accurate in predicting with cheurn.
- Recall: The proportion of accurately anticipated positive observations to all the actual class's observations is known as recall. In this instance 32% of the model has classified as no cheurn outcome and 91% as accurate with cheurn outcome .
- F1 score : F1 score is the weighted harmonic mean of precision and recall. The macro average of the f1 score is 39% and the weighted average is 45%. The accuracy is 40%.

```
Model Name: GaussianNB()

Best Parameters: {}

 [[182 388]
 [  9  88]]

              precision    recall  f1-score   support

   Outcome 0       0.95      0.32      0.48       570
   Outcome 1       0.18      0.91      0.31        97

    accuracy                           0.40       667
   macro avg       0.57      0.61      0.39       667
weighted avg       0.84      0.40      0.45       667
```

ROC Curve



Ensemble Model Results

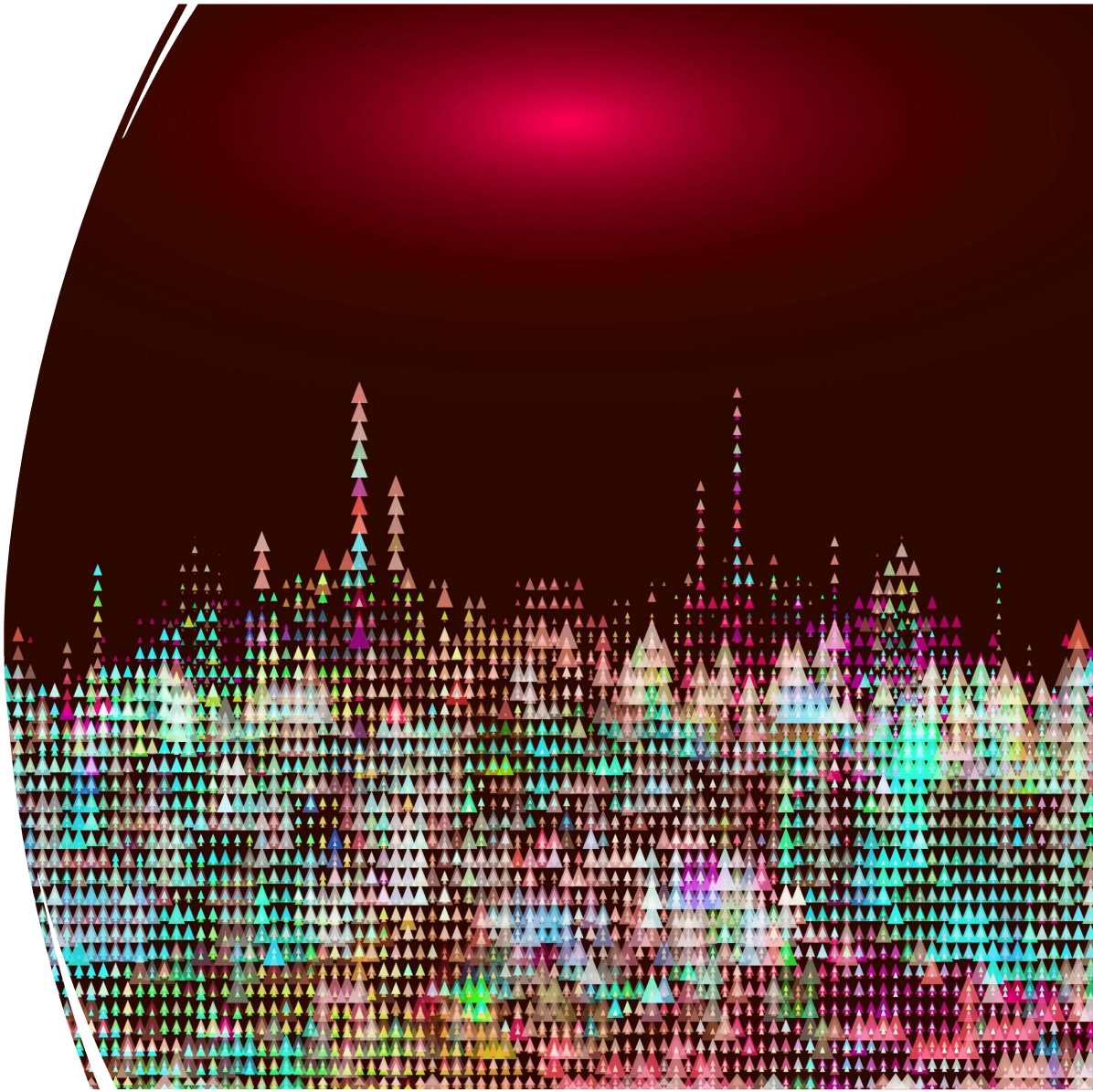
- Reasons to choose this model : we have both categorical and continuous data , and this model helps us in giving better results
- It offers better recall value results when compared to the logistical regression and naïve bayes model, with 84% accuracy as opposed to 83% and 45% provided by the other two methods, respectively.

Recommendation

- I would recommend logistical regression model because of the better recall score the model has given.

- More data set : In order to improve model accuracy we could get more observation into consideration. More data will help in an optimized output

- It also increases the predictive power of the algorithms by selecting the most crucial variables and eliminating the unnecessary ones. I advise Mr. John to use feature engineering to eliminate superfluous features and add new features to increase the predictive power of the algorithms.



Reference

Fatma Tetikoglu Week 1 to 13 - Data2204-Notes