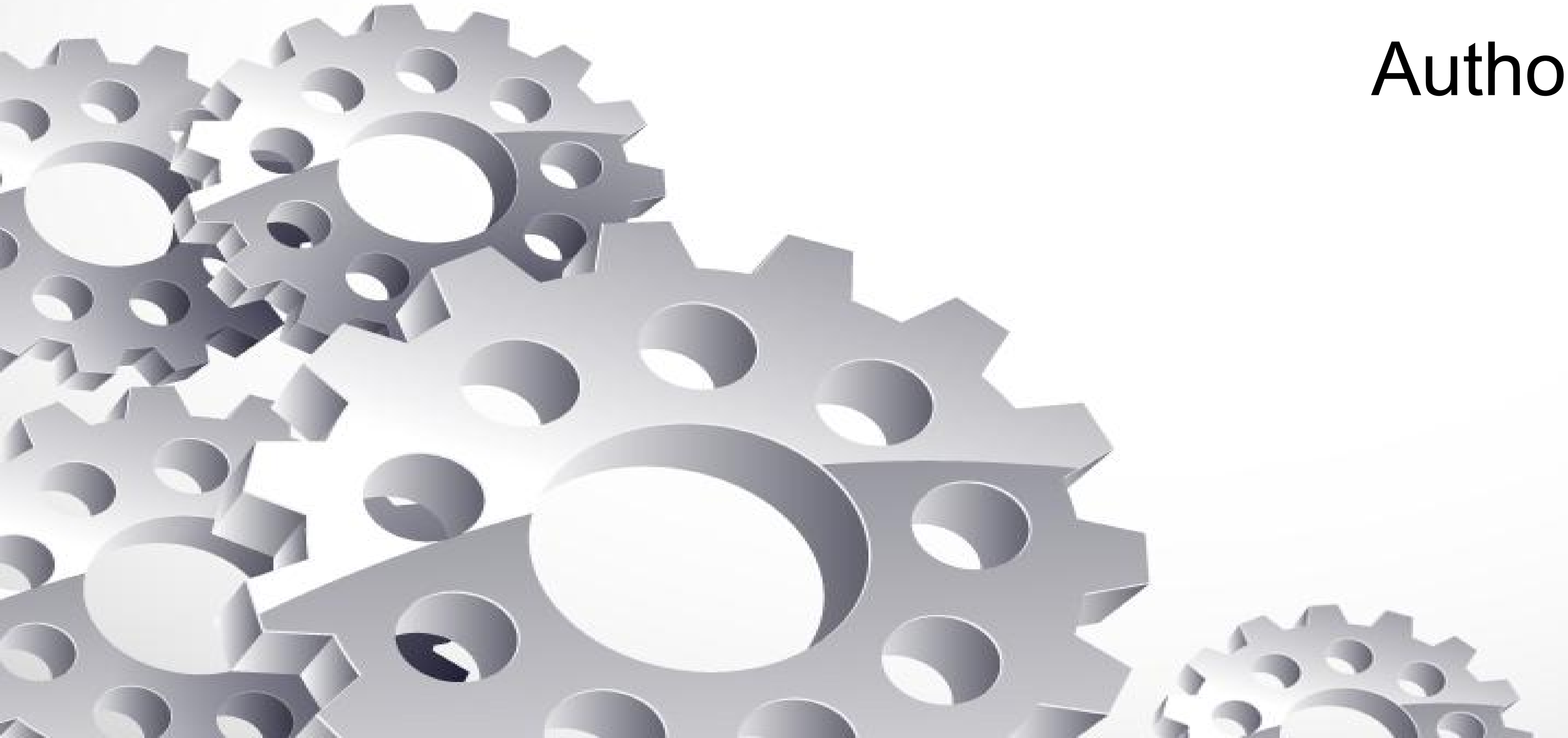
Data Science Phase 3 Project

Author: Nobert Akwir



Business Understanding.

Churn rate, in its broadest sense, is a measure of the number of individuals or items moving out of a collective group over a specific period. It is most commonly exppressed as the percentage of service subscribers who discontinue their subscriptions within a given time period. A high churn rate could adversely affect profits and impede growth.

Churn rate is an important factor in the tellecommunications industry. The mobile Telecommunication industry is becoming more saturated, with more and more customers swapping their registered services between competing companies.The churn rate not only includes when customers switch carriers but also includes when customers terminate service without switching. This measurement is most valuable in subscriber-based businesses in which subscription fees comprise most of the revenues.

Therefore, companies like SyriaTel, have realized thet they should focus their marketing efforts on customer retention rather than customer acquisition. In fact, it is less profitable for providers to attract new customers than prevent current customers from quitting. Hence, providers are engaged more and more in building predictive models in order to identify which customers are most likely to leave or churn, so that they offer them promotions to persuade them to keep using their lines.

Churn prediction is a management science problem for which a machine learning approach can be adopted. Based on Historical data, a model can be trained to classify customers as future chunners or non-chunners. Research shows that customer's behavioural features such as calls duration, calls count, refill/bill amount etc. , are good predictiors of churn.

Data Understanding.

Our dataset is based on historical customer churn behaviour for SriaTel, a tellecomunication company. The churn feature is our target variable and the remaining columns being our independent variable. Not more infor mation is provided on the dataset but the variable names provided are self explanatory in some sense. As with most real-world datasets, the column names are not perfectly described, hence we’ll have to do some research or use our best judgement if we have questions about what the data means.

Methodology

We are trying to find the best classification model too assist us predict our customer churn behaviour in the future. We will use an iterative approach to modelling where we build multiple models and try to see which one applies too our business problem the best using different appropriate evaluation metrics.

Modelling

For our iterative approach, we apply a logistic regression model, tune it to try see which one of it will be best. First we use a base model where we only encode the categorical values, then we scale the numerical values for our second model and finally penalize our model using the Lasso method. We also apply a KNN classifier and a Naive Bayes model.

Evaluation

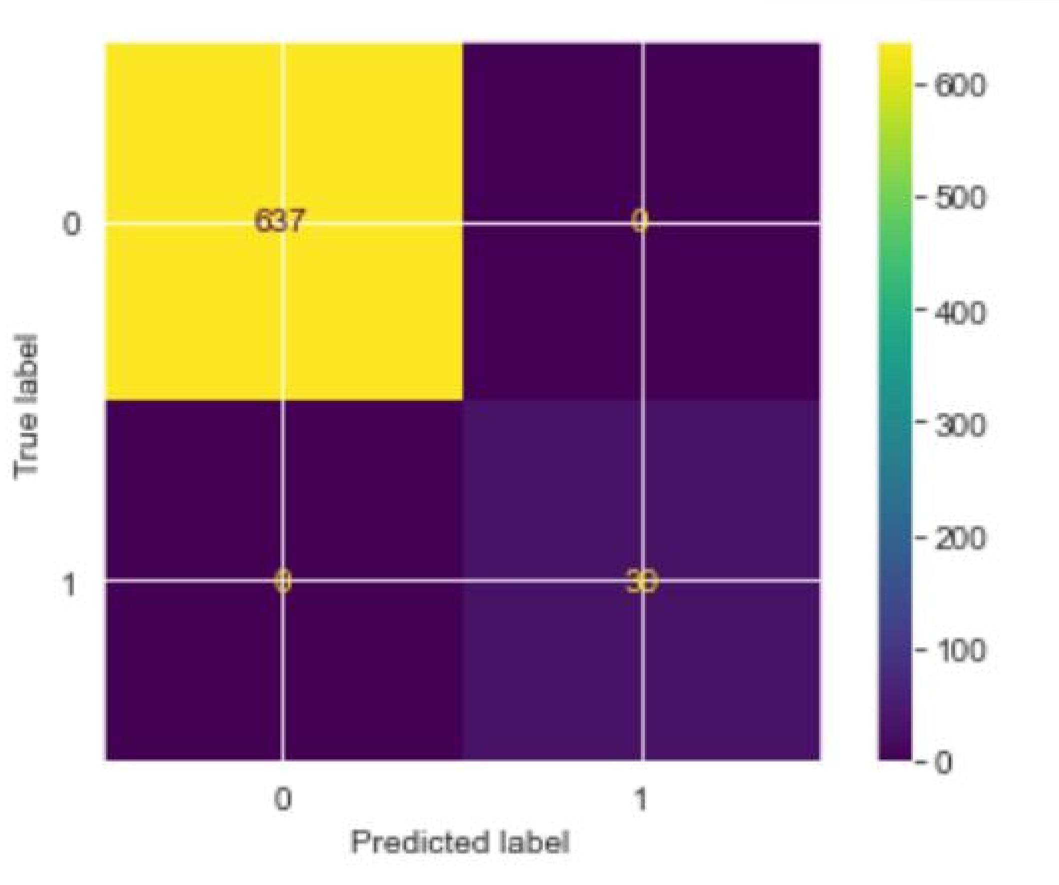
Base Model Results.

base_accuracy_score: 0.8545727136431784

base_precision_score: 0.5666666666666667

base_f1_score: 0.2595419847328244

base_recall_score: 0.16831683168316833



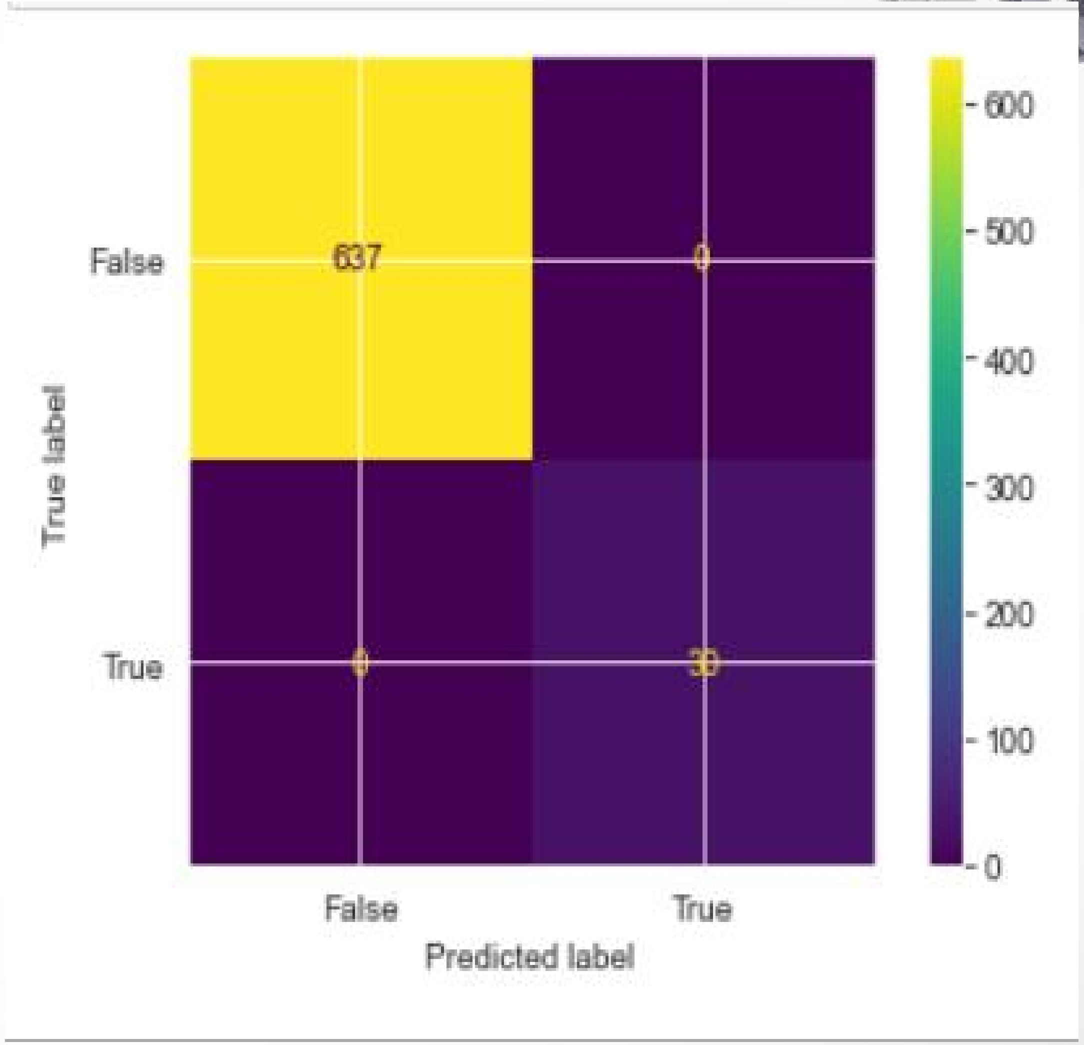
Nomalized features Model Results:

base_accuracy_score: 0.8545727136431784

base_precision_score: 0.5666666666666667

base_f1_score: 0.2595419847328244

base_recall_score: 0.16831683168316833



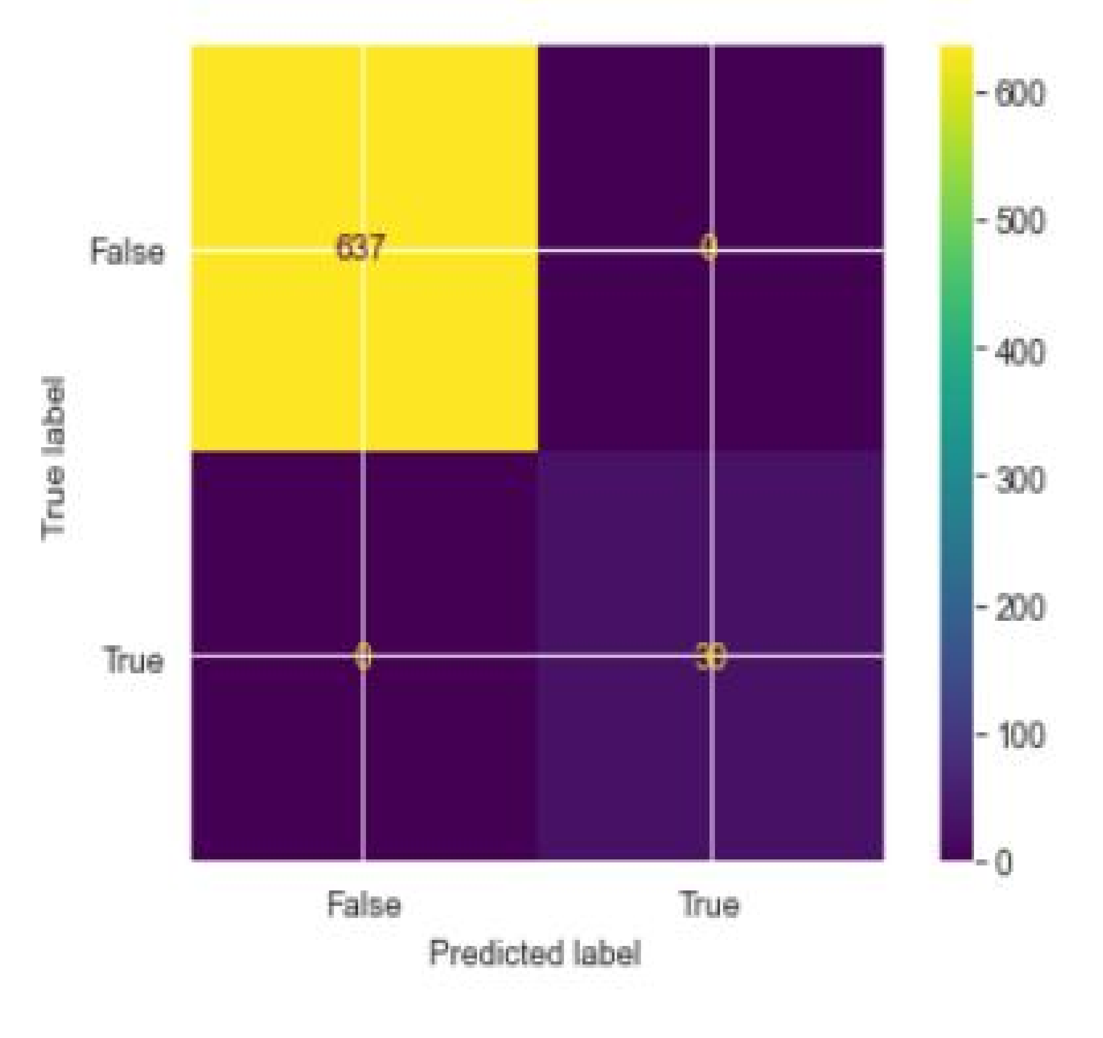
Penalized Model(Lasso Regression):

lasso_accuracy_score: 0.8545727136431784

lasso_precision_score: 0.5666666666666667

lasso_f1_score: 0.2595419847328244

lasso_recall_score: 0.16831683168316833



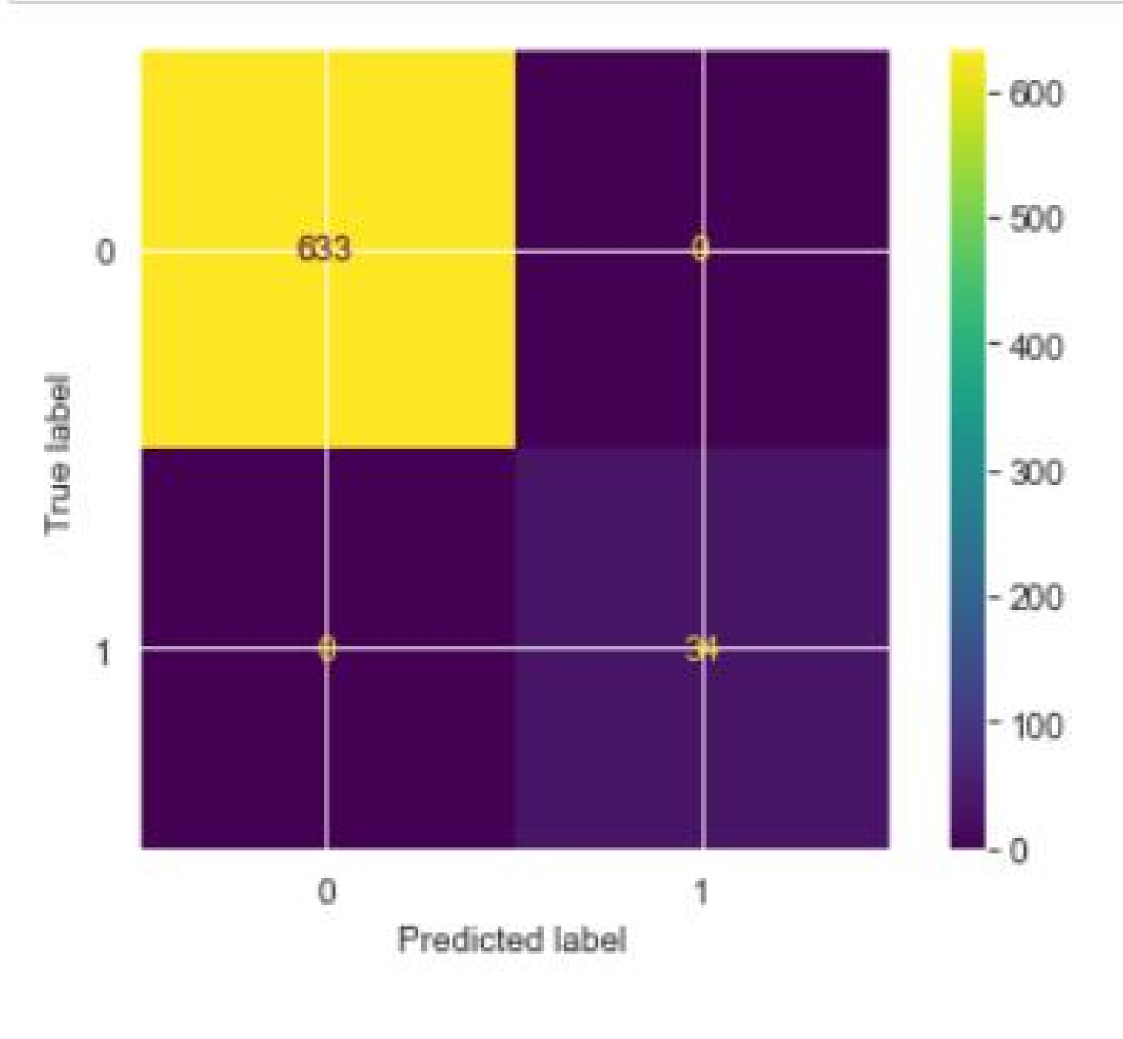
KNN Classifier Model:

knn_accuracy_score: 0.8755622188905547

knn_precision_score: 0.7647058823529411

knn_f1_score: 0.3851851851851852

knn_recall_score: 0.2574257425742574 3



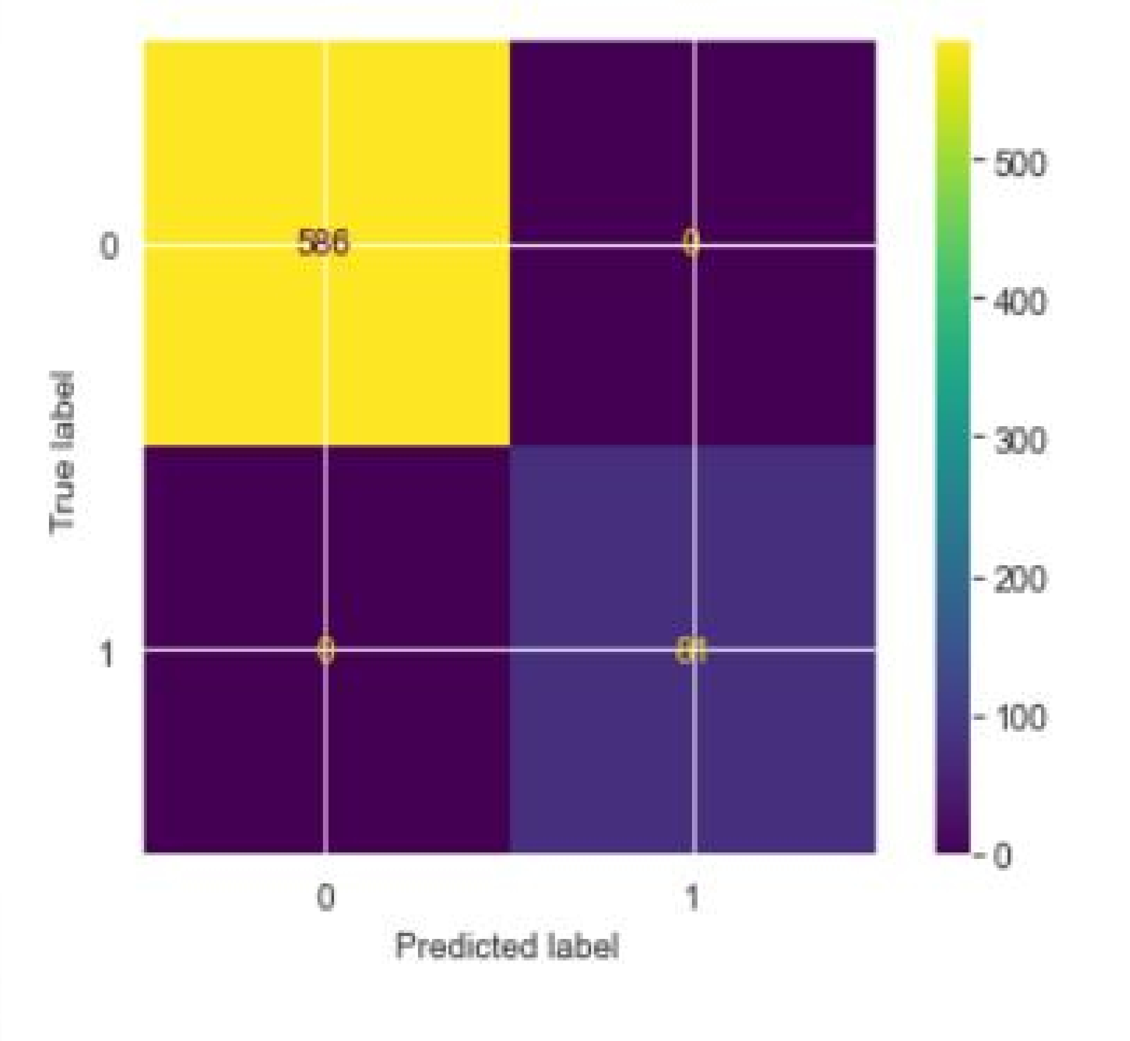
Naive Bayes Model:

naive_accuracy_score: 0.8440779610194903

naive_precision_score: 0.48148148148148145

naive_f1_score: 0.42857142857142855

naive_recall_score: 0.38613861386138615



From above results, we note that our three Logistic regression models perform the same and really poorly. The KNN Classifier performs slightly better than the Logistic Regression models and this is due to its acxuracy score and f1_score.

The Naive Bayes Model has a slightly lower accuracy score than the KNN Classifier but a better f1_score indicating slightly higher values on the pecision and recall scores to. Therefore the Naive Bayes model is the best model to adopt going foward before other models are applied to our dataset.

Reccomendations.

- For future model building, the Telco company could provide more data.
- Other classification algorithms should be applied to try predict the churn rate since the one’ applied currently haven’t really performed very well.
- There should be more metadata on the data so as to help undertand each variable.