Name: Sayyam Anil Gada
Student ID: 801307762

Churn prediction in telecommunication industry using kernel Support Vector Machines

Public Library of Science
https://doi.org/10.1371/journal.pone.0267935
Date: May 24, 2022

Nguyen Nhu Y, Tran Van Ly, Dao Vu Truong Son

Problem Statement

In the Telecommunication Industry, customer churn detection is one of the most important research topics that the company has to deal with retaining on-hand customers. Churn means the loss of customers due to existing offers of the competitors or maybe due to network issues.

The problem of churn prediction is to split the following tasks:

1. Input data analysis and pre-processing
2. Building a model to classify whether a customer will stay or churn
3. Evaluation of model concerning chosen metrics

Motivation

- In this age of fierce competitions, customer retention is one of the most important tasks for many companies.
- Churn rate has a substantial impact on the lifetime value of the customer because it affects the future revenue of the company and also the length of service.
- Due to a direct effect on the income of the industry, the companies are looking for a model that can predict customer churn.
- In this project, I will create a classification machine learning model which will predict whether a customer will be retained or churned.

Data

Telecom Customer Churn Dataset

- The Customer Churn dataset contains information on 3333 customers from a French based telecom company named Orange Telecom.
- Each record represents one customer, and contains statistics about no of calls, minutes,etc
- Download Link: https://www.kaggle.com/datasets/mnassrib/telecom-churn-datasets

Survey on Related Work

Title: Churn Prediction of Customer in Telecom Industry using Machine Learning Algorithms

Authors: V. Kavitha, S. V Mohan Kumar, G. Hemanth Kumar, M. Harish

International Journal of Engineering Research & Technology (IJERT), Vol. 9 Issue 05, May-2020

URL: https://www.researchgate.net/publication/341870705_Churn_Prediction_of_Customer_in_Telecom_Industry_using_Machine_Learning_Algorithms

The paper does comparison of classification and prediction of telecom customer churn on three different algorithms-

- Random Forest
- Logistic Regression
- eXtreme Gradient Boosting

The paper than compares statistics of each algorithm. Accuracy of each algorithm is close but Random Forest comes on top having 80% accuracy followed by Logistic Regression and eXtreme Gradient Boosting having 79% and 78% accuracies.

Survey on Related Work

Title: Churn Prediction Using Machine Learning and Recommendations Plans for Telecoms

Authors: Khulood Ebrah1, Selma Elnasir

Scientific Research Publishing Journal of Computer and Communications, 2019, 7, 33-53

DOI: https://doi.org/10.4236/jcc.2019.711003

The paper does visualization of the dataset along with comparison of classification and prediction of telecom customer churn on three different algorithms-

- Naïve Bayes
- Support Vector Machine
- Decision Tree

The paper than explains the models performance which is measured by area under curve where the best

AUCs are predicted by Support Vector Machine.

Summary of the Method

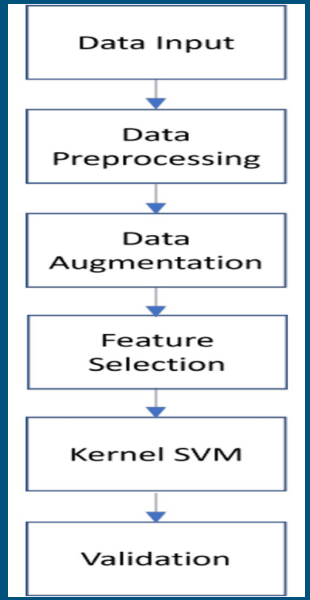
In the research paper “Churn prediction in telecommunication industry using kernel Support Vector Machines” the authors have proposed a kernel Support Vector Machines algorithm based classification model which will predict whether a customer will churn or not. Dimension reduction strategies such as Sequential Forward Selection (SFS) and Sequential Backward Selection (SBS) are applied to the dataset to find out the most important features. The model has an accuracy of 98.9%

Why does it make sense?

- Support Vector Machines is effective in high dimensional spaces.
- SVM uses a subset of training points in the decision function (called support vectors), so it is also memory efficient.
- Versatile: different Kernel functions can be specified for the decision function. Common kernels are provided like linear, polynomial, rbf and sigmoid, but it is also possible to specify custom kernels.
- The paper finds out that rbf kernel SVM is the most effective model of classification among other classification algorithms and kernels.

Source:
https://scikit-learn.org/stable/modules/svm.html

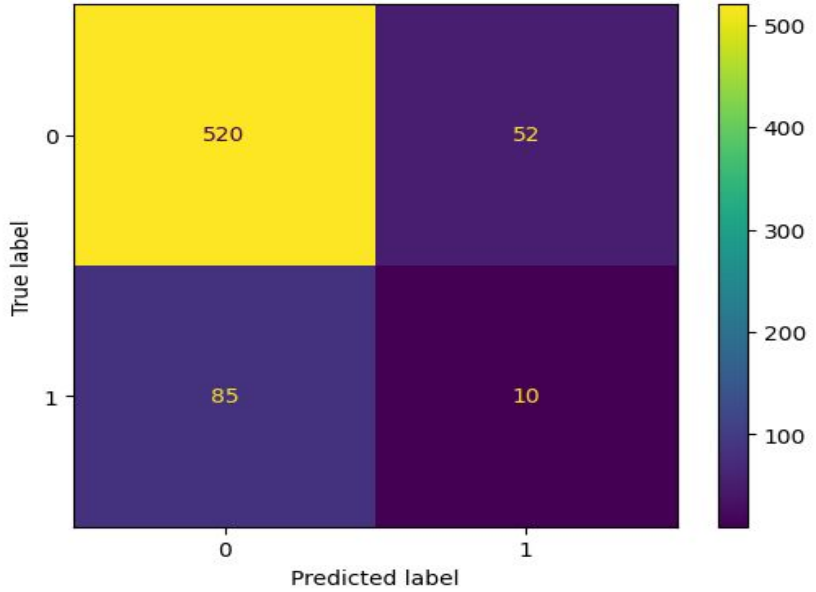
Given Below is the Diagrammatic representation of the Support Vector Machines model used in the research paper:



Results- Sigmoid

Train Accuracy: 0.794074
Test Accuracy: 0.794603
MCC: 0.017276718360329745

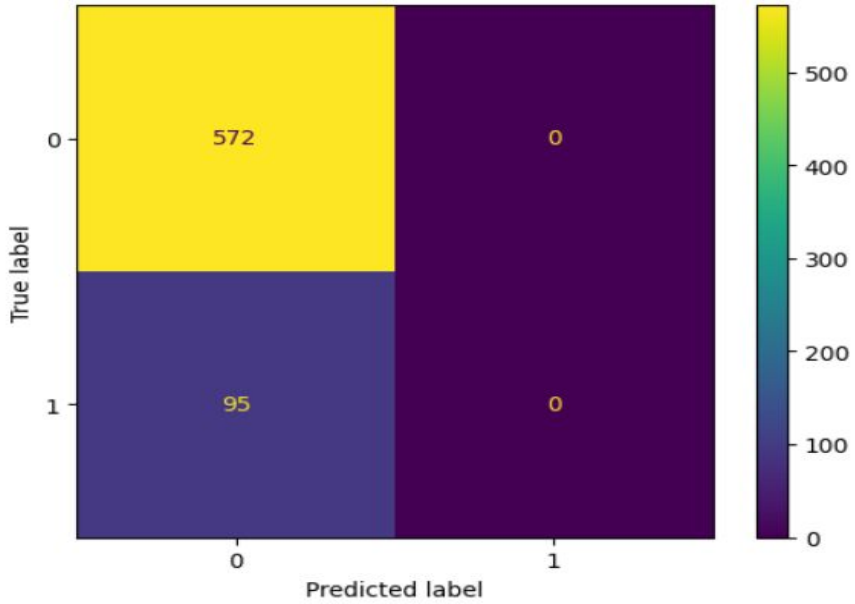
	precision	recall	f1-score	support
0	0.86	0.91	0.88	572
1	0.16	0.11	0.13	95
accuracy			0.79	667
macro avg	0.51	0.51	0.51	667
weighted avg	0.76	0.79	0.78	667



Results- Linear

Train Accuracy: 0.854464
Test Accuracy: 0.857571
MCC: 0.0

	precision	recall	f1-score	support
0	0.86	1.00	0.92	572
1	0.00	0.00	0.00	95
accuracy			0.86	667
macro avg	0.43	0.50	0.46	667
weighted avg	0.74	0.86	0.79	667

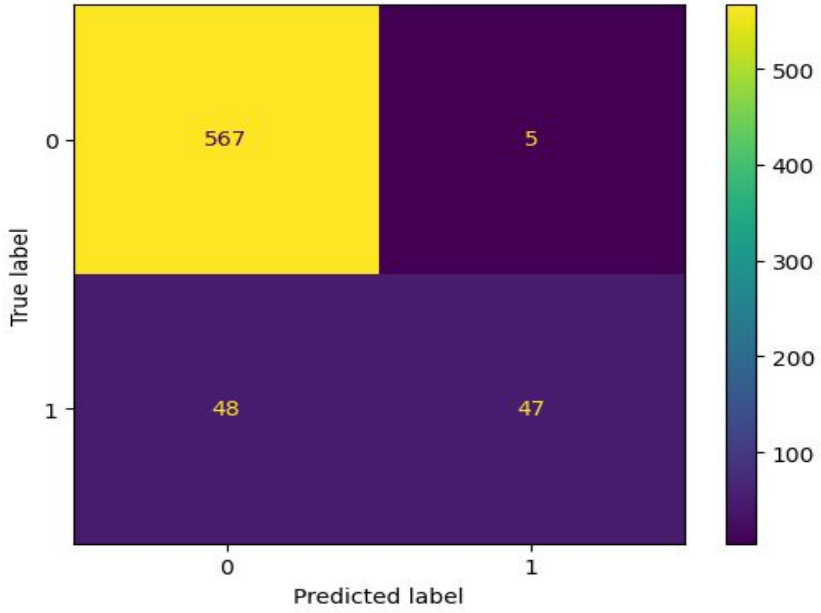


Results- rbf

```
Train Accuracy: 0.943736
Test Accuracy: 0.920540
MCC: 0.6335091761022514
              precision    recall  f1-score   support

           0       0.92      0.99      0.96       572
           1       0.90      0.49      0.64        95

    accuracy                           0.92       667
   macro avg       0.91      0.74      0.80       667
weighted avg       0.92      0.92      0.91       667
```

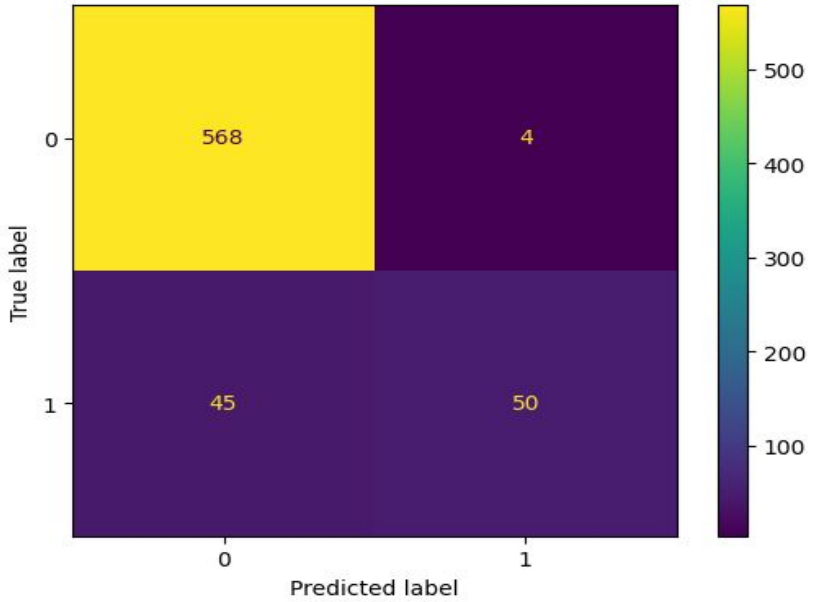


Results- Polynomial

```
Train Accuracy: 0.942236
Test Accuracy: 0.926537
MCC: 0.6653805094313474
              precision    recall  f1-score   support

           0       0.93      0.99      0.96       572
           1       0.93      0.53      0.67        95

    accuracy                           0.93       667
   macro avg       0.93      0.76      0.81       667
weighted avg       0.93      0.93      0.92       667
```

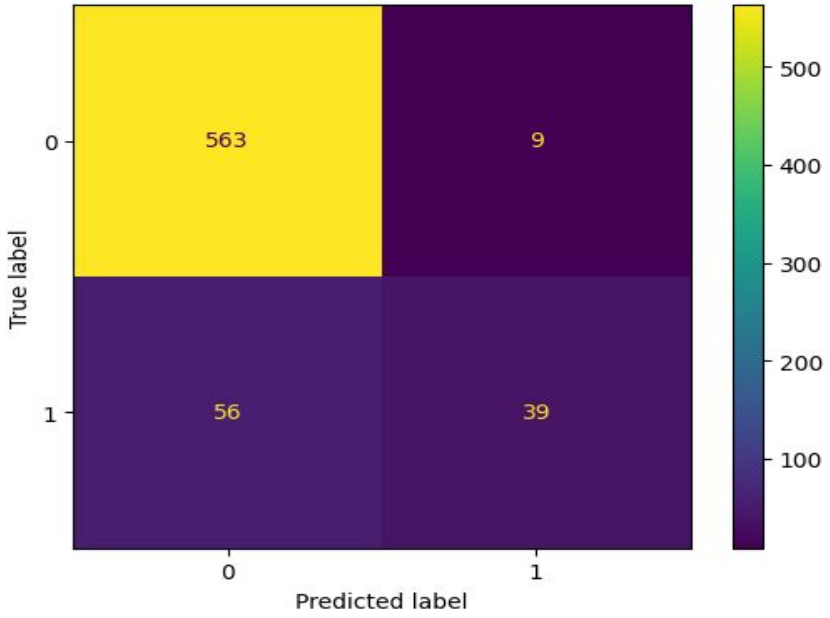


Results

```
Train Accuracy: 0.887097
Test Accuracy: 0.902549
MCC: 0.5339028910261884
              precision    recall  f1-score   support

           0       0.91      0.98      0.95       572
           1       0.81      0.41      0.55        95

    accuracy                           0.90       667
   macro avg       0.86      0.70      0.75       667
weighted avg       0.90      0.90      0.89       667
```



Conclusion

- rbf has best train accuracy
- Polynomial performs the best
- Feature Selection