**Assignment 02**

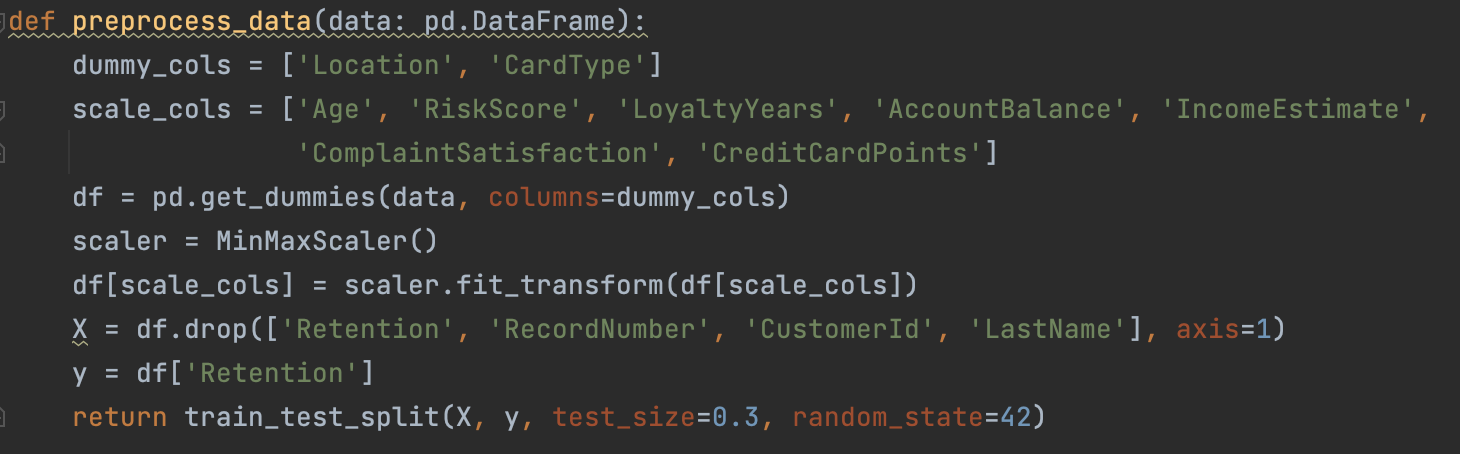
1-2: Data Exploration and Preprocessing:

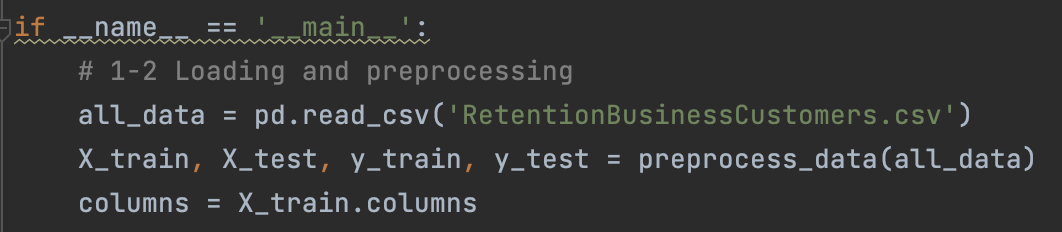
Looking at the data we decided to make some changes so that it will be easier to work with. First, we figure that the “location” and “CardType” feature may have a significant effect on retention but they were currently just categories and not numbered and so we decide to turn them into dummy variable, as machine learning algorithms typically work with numerical representations.

Secondly, we decided to scale quite a few of the features so that they will all be in the same measurement unit, originally they range of numbers was very different for each feature and so to make our models more interoperable we decided to scale them.

Thirdly, we chose to drop the feature we were quite confident would have no direct effect on our model’s results, such as RecordNumber, CustomreID and the customer’s last name.

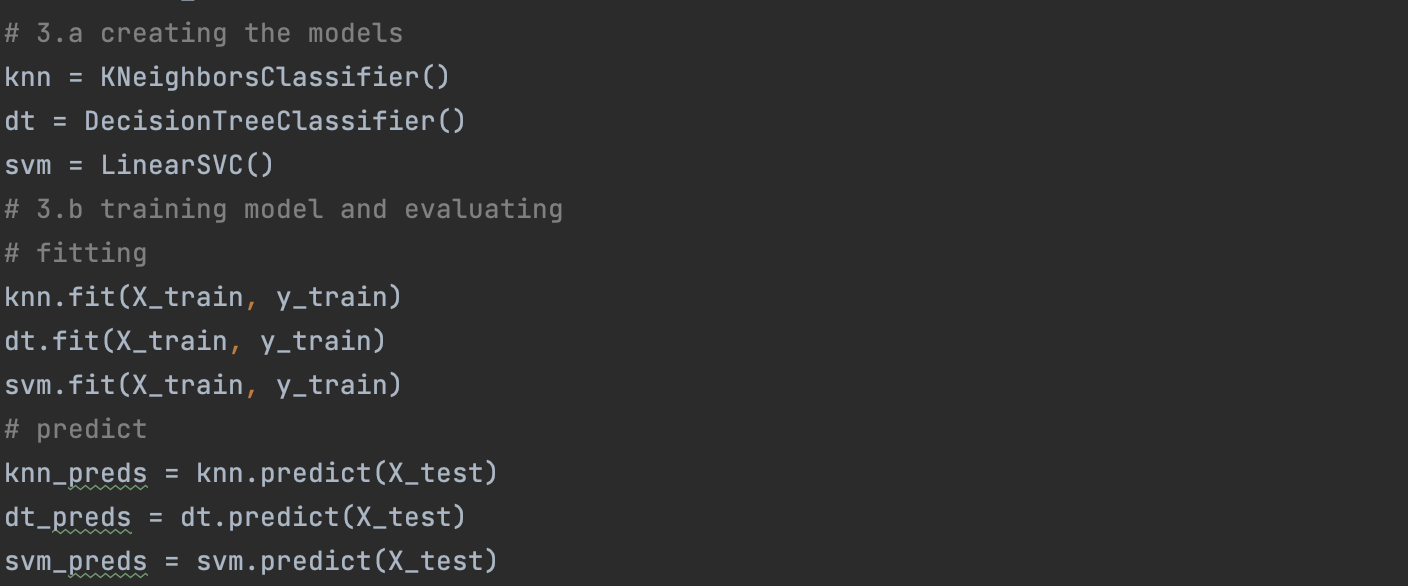
Finally, we split our data into X and y and split it into train and test sets, 70:30.af

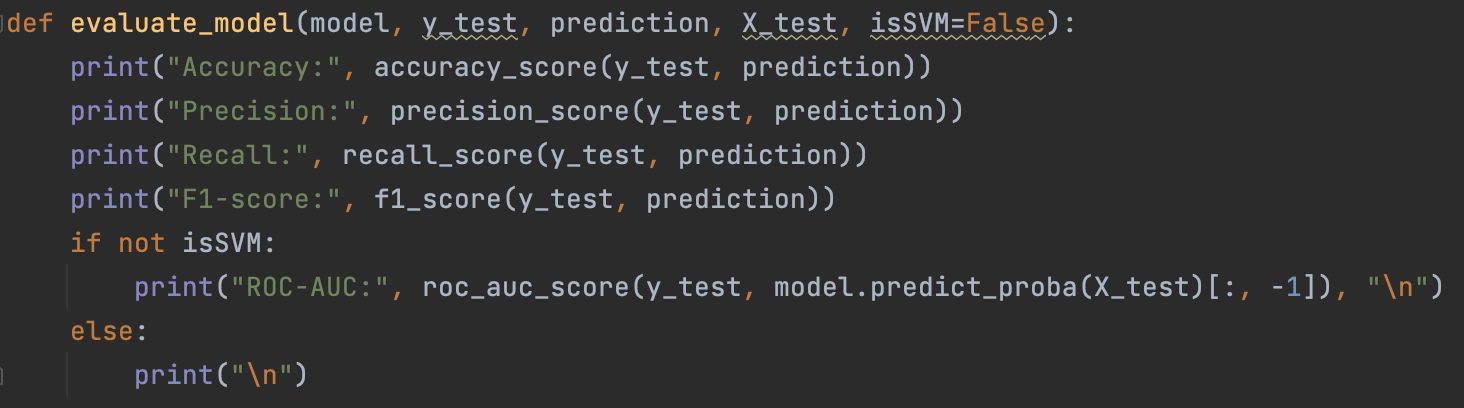


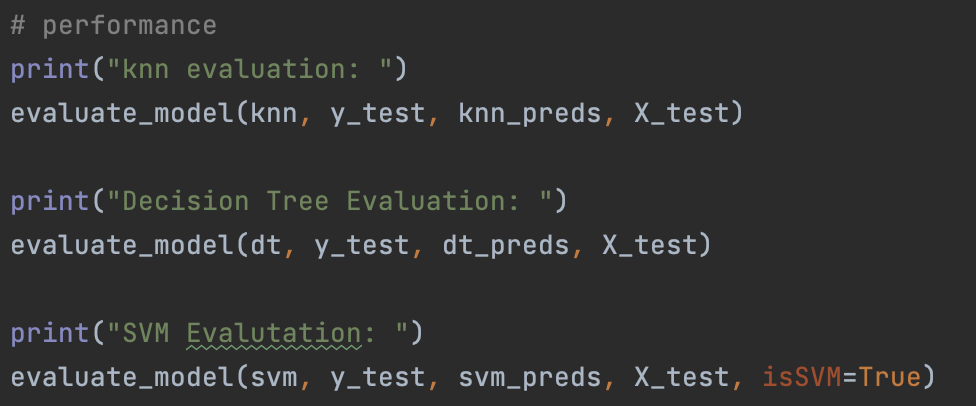


3: Model Training and Evaluation

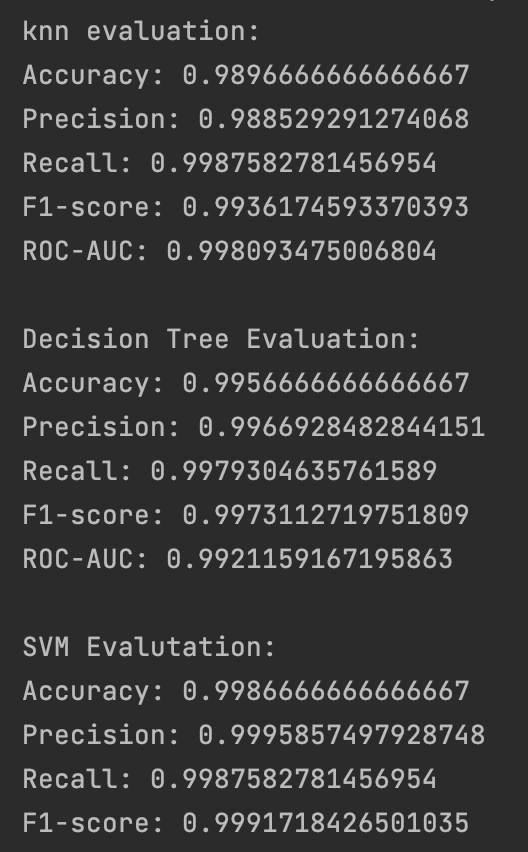
The models we chose to train and evaluate are: KNN, Decision Tree and Linear SVM







The results were:



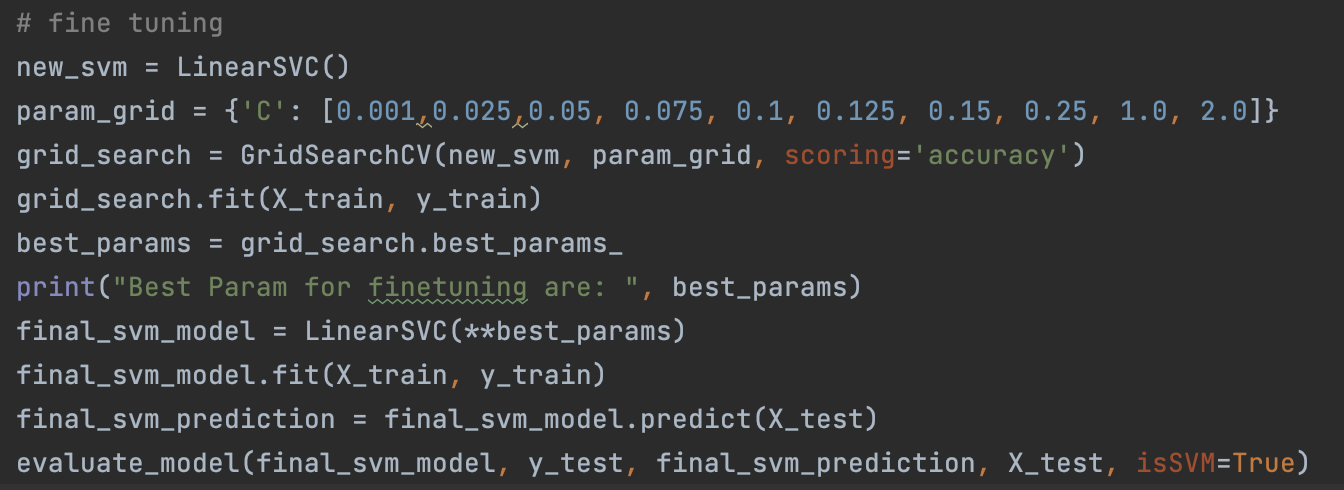
As you can see all the models did really well (almost 100% in all evaluations) but if we have to choose a winner it is SVM. From the results it is hard to see the strength and weaknesses of each model, as they all did really well. So, we will do a general analysis of the models.

It is important to add that ROC-AUC score is not possible to calculate for Linear-SVM as it doesn’t return probabilities and so we cannot calculate the ROC-AUC curve (there is no built in function in sklearn).

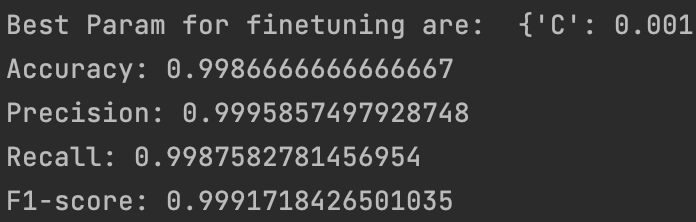
|  |  |  |  |
| --- | --- | --- | --- |
| Model | KNN | Decision Tree | Linear SVM |
| Strengths | * Intuitive and simple to understand. * No assumptions about data distribution. * Suitable for multi-class classification. | * Easy to interpret and visualize. * Handles numerical and categorical features. * Automatically captures nonlinear relationships. | * Effective in high-dimensional spaces. * Performs well with small sample sizes. * Robust to overfitting with appropriate regularization. |
| Weakness | * Computationally expensive during prediction. * Sensitive to irrelevant features. * Requires appropriate feature scaling. | * Prone to overfitting. * Sensitive to small data changes. * Struggles with unbalanced data. | * Computationally expensive for large datasets. * Sensitivity to hyperparameter tuning. * Difficult to interpret individual feature significance. |

4: Model Optimazation

We chose to optimize our SVM model using grid search and find the best C



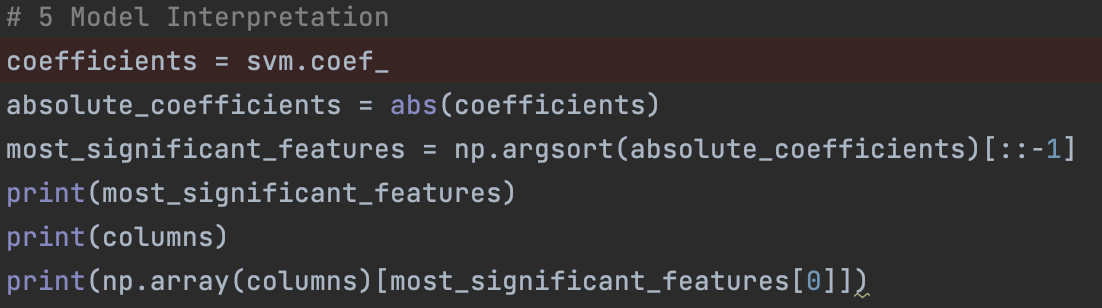
The results were:



As you the fine-tuned model resulted the same as previously (which was really good)

5: Model Interpretation

We chose to analyze the coefficients.



The order of the coefficients is (from smallest to largest):

A screen shot of a computer

Description automatically generated with low confidence

Furthermore, looking at the coefficient values “HasComplaint” is significantly larger than the other. Therefore, we recommend looking closely into any customer that has complained because the chance of them no being retained in the bank is the largest. Next for some reason the customer’s location also has a significant effect on his retention which should also raise a red flag.

6: Conclusion

In conclusion we tested three different models, KNN, DT and SVM, each with its own unique strengths and weaknesses. All models showed extremely good results after they were fitted with the training data and so it was hard to show through the result their differences.

We then chose to optimize our “best” mode (SVM) with GridSearch, this method tests different hyperparameters (in case of SVM C) and compares the model if it were to be fitted with each of the hyperparameters. In this case the best parameter was C=0.001 but our new model with the new hyperparameter showed the same results as previously (which again were almost perfect).

Lastly, we analyzed our fitted models’ coefficients and ordered them by absolute value in order to conclude which of the had the largest effect on the model’s prediction.

The results showed that Has Complaint has the largest effect on customer retention (which makes sense that a customer who has complained will not stick), then if the customers location is Israel (interesting) and more. It is important to add that HasComplaint’s coefficient was remarkably larger than the rest.