BIS 634 Final Report:

Bank Customer Churn Prediction

Ji Yoon Lee

BIS 634 Computational Methods for Informatics

December 21, 2022

Introduction

Background

Churn prediction is the prediction of which customers are likely to stop subscriptions, products, or services with a business. Customer churn prediction is a crucial task for all business. In banks, customer churn prediction can help management target those customers that are likely to churn with promotions. Another benefit of churn prediction is that it provides insights into which factors are important to retain customers. Many businesses have begun to develop models to accurately predict customer churn. Using the large amount of customer data and machine learning algorithms, businesses are developing their own churn prediction models.[1] This report uses current methods to analyze and predict bank customer churn and introduces a web interface to visualize these results.

Data Resources and FAIRness

The dataset was downloaded as a .csv format from Kaggle, a public data repository. The dataset is owned by Gaurav Topre and contains no specific information about license.[2] The dataset follows the data FAIRnesss guideline.[3]

- Findability: The data and metadata can easily be found by everyone on Kaggle through 'Bank Customer Churn Dataset'.
- Accessibility: The data and metadata are open and free from a public repository - no permission is needed to download the data from Kaggle.
- Interoperability: The data is stored in a .csv format.
- Reusability: The data and metadata are well annotated. However, the license information is not specified.

Analyses Questions

The two analyses questions that will be addressed in this report are:

- Which features (numerical and categorical) have an effect on churn?
- Which model most accurately predicts customer churn?

The first question is important because it helps identify which factors are related to whether a customer churned or not. For example, if female customers have a higher churn rate than male customers, then the business can increase products or services targeted for female customers. The question will be answered by analyzing the differences in distribution of features between churned and not churned customers.

The second question takes the analysis from the first question a step further – it develops a model that will accurately predict which customers are likely to churn. The prediction model can be used to identify which customers are likely to churn and management can target those specific customers with special promotions or emails.

Exploratory Data Analysis

Data Summary

The 'Bank Customer Churn Dataset' contains 10000 entries and 12 variables. The variables are[2]

- customer_id: account Number [numerical value]
- credit_score: credit score [numerical value]
- country: country of residence [categorical value: France, Spain, Germany]
- gender: sex of the customer [categorical value: Male, Female]
- age: age of the customer [numerical value: years]

- tenure: number of years customer owned the bank account in ABC Bank [numerical value: years]
- balance: account balance [numerical value: euros]
- products number: number of products from the bank [numerical value of 1 to 4]
- credit card: whether customer owns a credit card [1: Yes, 0: No]
- active member: whether customer is an active member of the bank [1: Yes, 0: No]
- estimated salary: salary of account holder [numerical value: euros]
- churn: churn status [1: churned, 0: not churned]

Data Preprocessing and Cleaning

The dataset was clean; it contained no missing value and no duplicate rows. There was not much data cleaning to perform. The values of the dummy variables credit_card and active_member was replaced with “Yes” or “No”, and the values of variable churn were replaced with “churned” and “not churned”. The column customer_id was dropped because it did not contain any meaningful information. The variables tenure and products number are numerical variables, but they contained only 10 and 4 unique values, so are considered categorical variables.

Univariate Analysis

The distribution of credit score is approximately normal with at 600 to 700. The distribution of age is skewed to the right – most of the customers are aged between 20 to 40. The balance distribution of balance is approximately normal with a peak at balance equal to 0. This could be an error but will not be removed because it might be associated with customer churn. The distribution of estimated salary is approximately uniform (Figure 1).

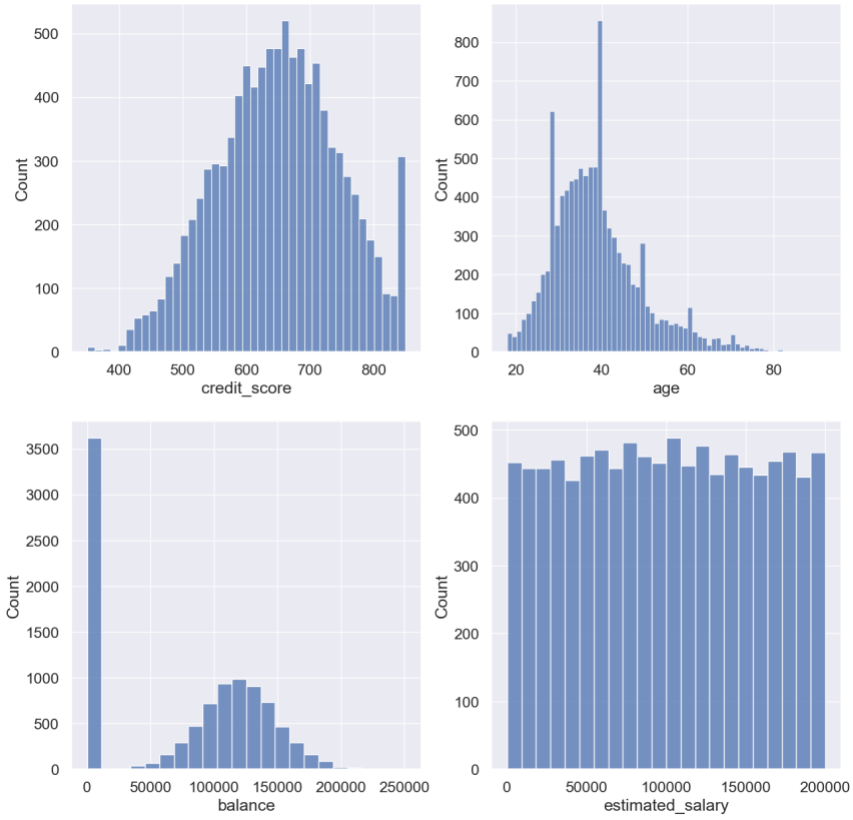

Figure 1: The Distribution of Continuous Variables

The customers are from 3 different countries – France, Spain, and Germany. There is a slightly larger number of male customers. Customers have owned their accounts from 0 to 10 years. Customers have 1 to 4 products in the bank (most have 1 or 2 products). There is a larger number of customers with credit card and an approximately equal number of active and not active members (Figure 2).

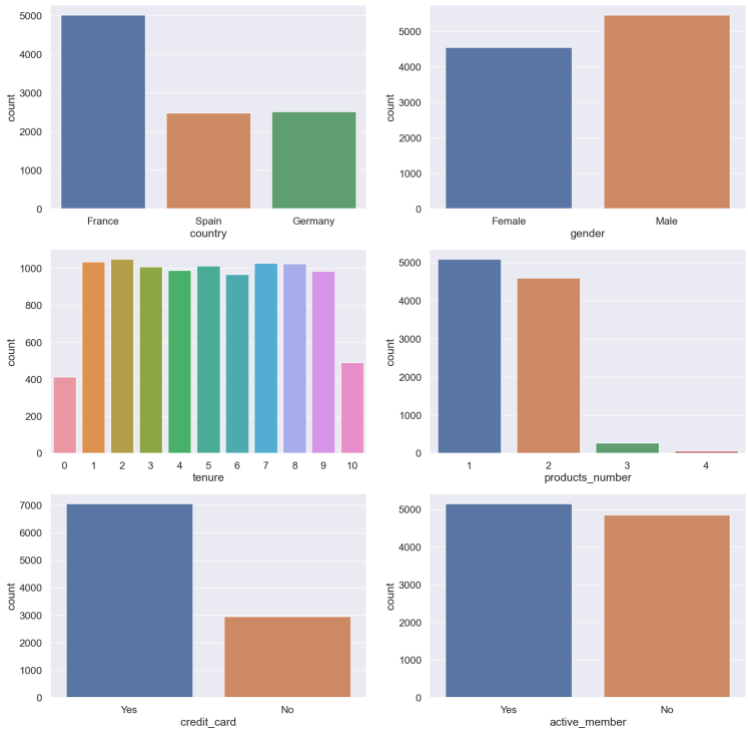

Figure 2: The Distribution of Categorical Variables

Approximately 80% of customers have churned (Figure 3). The dataset is imbalanced.

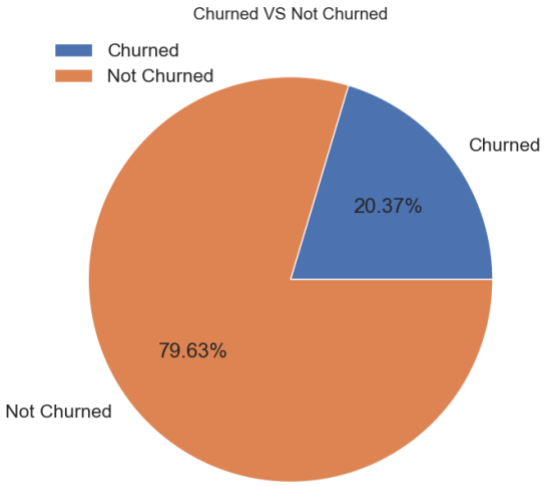

Figure 3: Pie Chart of Target Variable – Churn

Bivariate Analysis

Which categorical features have an effect on churn?

To account for the imbalanced number of churned and not churned customers, the proportions were normalized to churn category. Customers from France were less likely to churn, while customers from Germany had a higher churn rate. Females have churned more than males. Whether customers owned a credit card and number of years the customer had been with the bank had no effect on whether customers churned. Customers with one account were more likely to have churned and customers with 2 accounts were more likely to not have churned. Non active members churned with a higher proportion than active members (Figure 4).

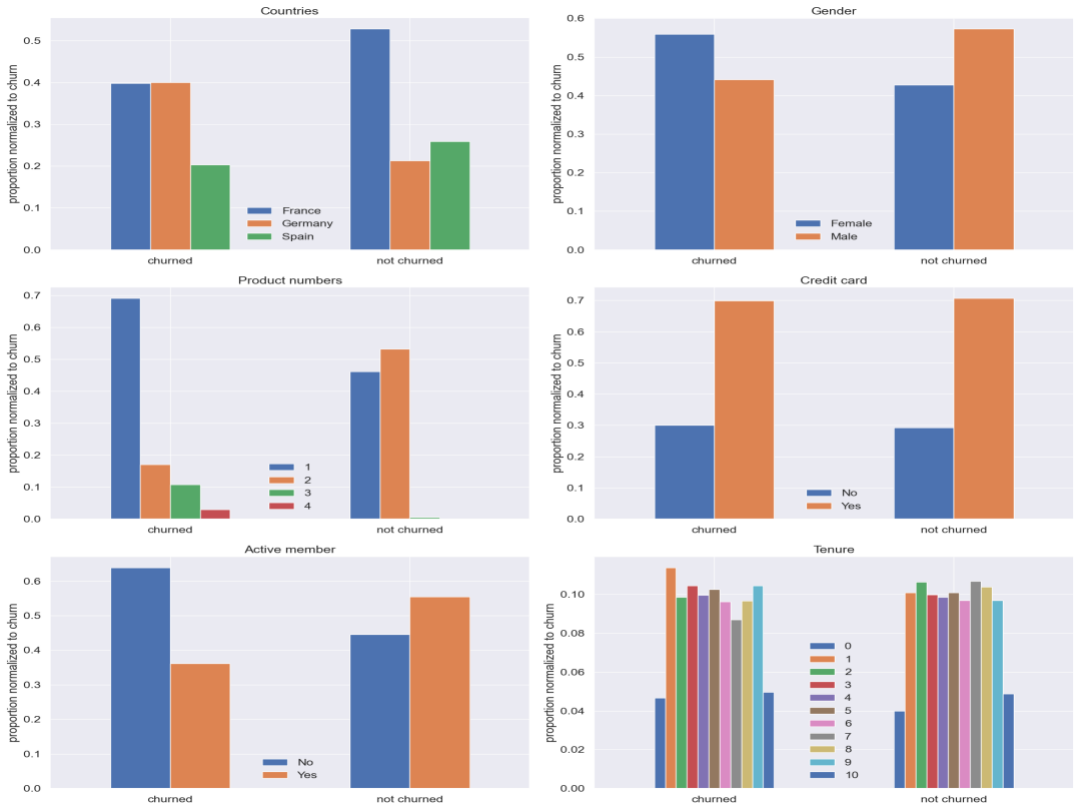

Figure 4: Distribution of Categorical Features in Respect to Churn

Which numerical features have an effect on churn?

There is no significant difference in credit score and estimated salary distribution between the churned and not churned group. The age of customers that have churned are higher than the age of customers that have not churned. The balance of customers that have churned are slightly higher than the balance of customers that have not churned (Figure 5).

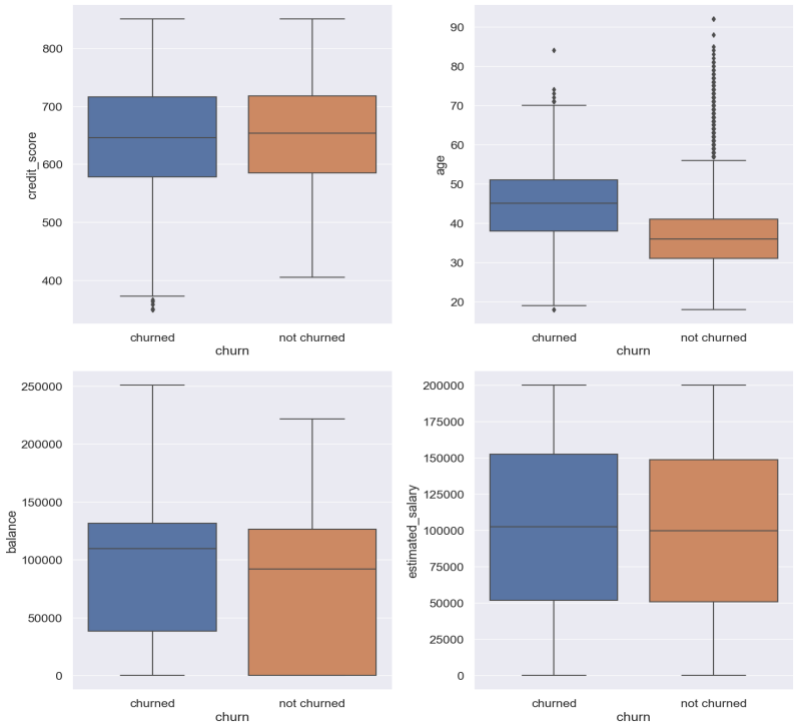

Figure 5: Distribution of Numerical Features in Respect to Churn

Age had a positive correlation with churn – older customers are more likely to churn than younger customers. There were no significant correlations between other predictor variables (Figure 6).

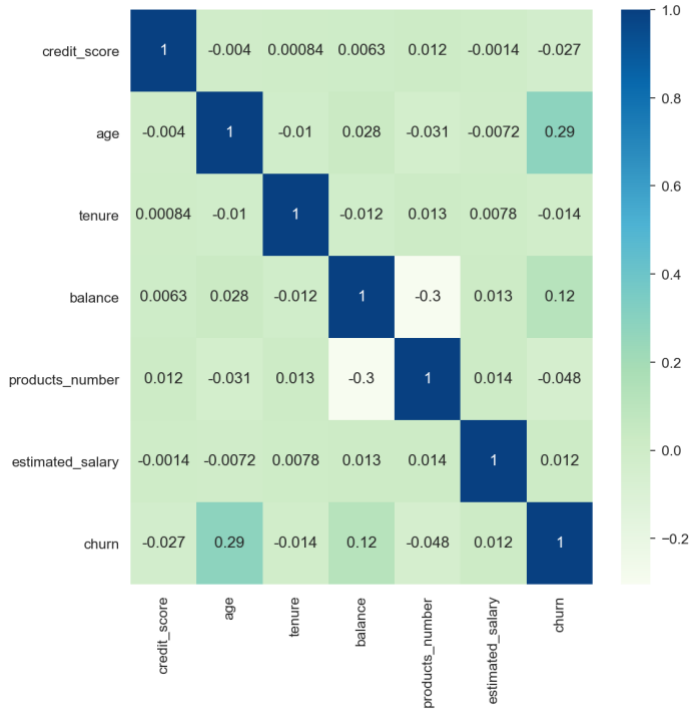

Figure 6: HeatMap of Variables

Machine Learning Models

Which model most accurately predicts customer churn?

K-Nearest Neighbors (KNN)

K-Nearest Neighbors (KNN) algorithm is a "non-parametric, supervised learning classifier, which uses proximity to make classifications or predictions about the grouping of an individual data point"[4]. The K in KNN is a parameter that indicates count of the nearest neighbors. It is important to find an optimal value of K to build a model that achieves maximum accuracy. To select the optimal K value, a plot of error rate with different K values was generated and the K value that corresponds to the minimum error rate is selected. For this dataset, the minimum error was 0.172 at K = 23 (Figure 7). Data was preprocessed before the model was implemented – the categorical variables were encoded using the LabelEncoder function, all predictors were standardized using the StandardScaler function, and the data was split into 70% train dataset and

30% test dataset. The KNN model was implemented using the KNeighborsClassifier function in the scikit-learn library. The accuracy of KNN model with K=23 was 82.77% (Figure 8).

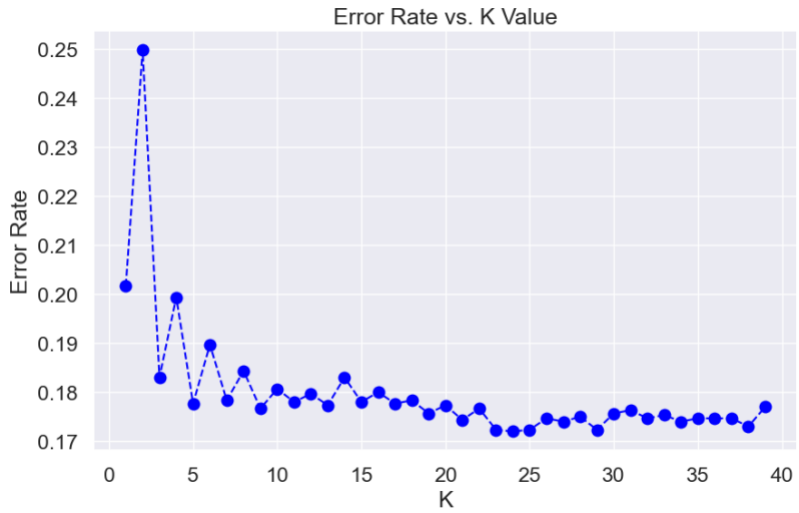

Figure 7 Error Rate vs. K Value

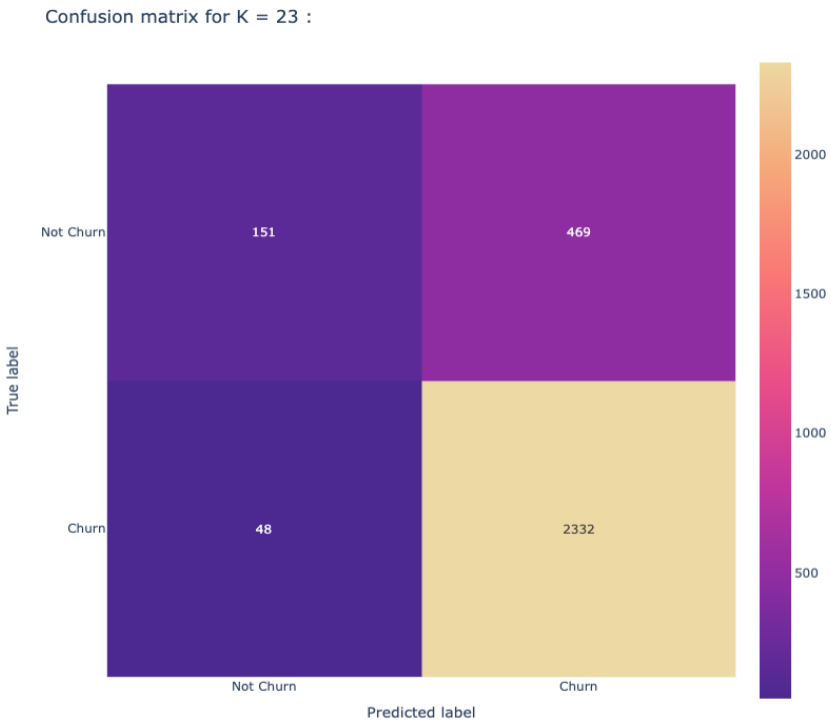

Figure 8 KNN Confusion Matrix for K=23

Support Vector Machine (SVM)

The Support Vector Machine (SVM) maps data to "high-dimensional feature space so that data points can be categorized".[6] The algorithm finds a hyperplane separator between the two classes.[6] The kernel function and the regularization parameter are the parameters of SVM. The GridSearchCV function from sklearn.model_selectionis was used to tune the hyper-parameter of the model. The SVM model achieved the highest accuracy when the kernel function was rbf and the regularization parameter C=1. The data was preprocessed the same way as KNN, and the SVM model was implemented using the SVC function in the scikit-learn library. The accuracy of SVM model with a rbf kernel and C=1 was 84.9% (Figure 8).

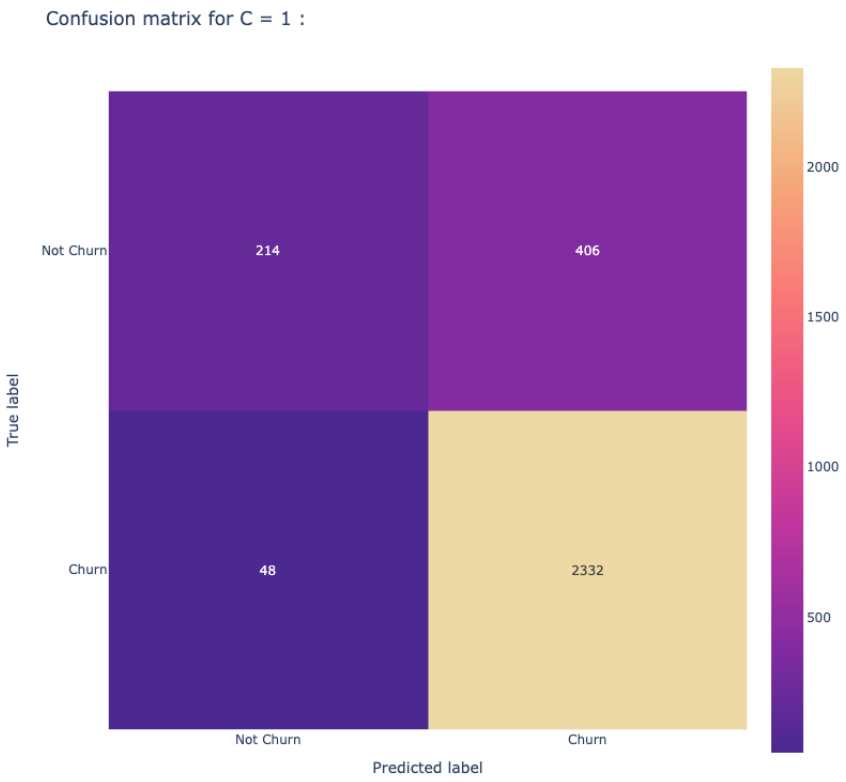

Figure 9 SVM Confusion Matrix for RBF Kernel and C=1

Server API and Web front-end

Backend API server

A Flask app was created. The app,py file contains the flask app, along with all the API routes and its corresponding html files (these files can be found under the templates folder). A separate models.py that contained the KNN model function and the SVM model function was created and imported into the app.py. The main route is directed to analyze_knn or analyze_svm depending on whether the user has chosen the KNN model or the SVM model. The parameters entered by the users are taken in to display the confusion matrix using the POST method. Entering export FLASK_APP=app and flask run on the terminal runs the app.

Web front-end

The web interface is a single page website. There are two options to select from: K-Nearest Neighbor and Support Vector Machine (SVM). When users select the K-Nearest Neighbor, users are asked to enter a K value. When users select the Support Vector Machine option, users are asked to enter a C value. Clicking the predict button navigates users to the result page. The result page for both options displays the accuracy score along with an interactive confusion matrix of the predicted and true labels for the output variable churn status.

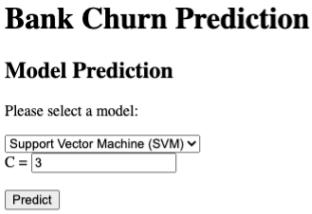

Figure 10 The Web Interface – Model Selection

Discussion

This report answers the two analyses questions and describes a web interface that allows users to select a model, specify the parameters, and visualize the results.

Which features have an effect on churn?

There were several factors that were associated with whether customers churned. From bivariate analysis, it was observed that customers from Germany, were Female, had only one product in the bank, non-active members, or old customers were more likely to churn (not joint effect but each category is a risk factor of churn).

Which model most accurately predicts customer churn?

The SVM model with a rbf kernel and a regularization parameter C=1 (no regularization) had the highest accuracy of 84.9%.

A folder that contains the data, template, and apps can be found on github.com/jiyoonlee96.

References

[1] Paddle.com. (n.d.). *Customer churn 101: What is it and why does it matter so much to SaaS businesses?* https://www.paddle.com/resources/customer-churn

[2] Topre, G. (n.d.). *Bank Customer Churn Dataset* [Dataset]. https://www.kaggle.com/datasets/gauravtopre/bank-customer-churn-dataset?datasetId=2445309&sortBy=voteCount

[3] GO FAIR initiative. (2022, January 21). *FAIR Principles*. GO FAIR. https://www.go-fair.org/fair-principles/

[4] *What is the k-nearest neighbors algorithm? | IBM*. (n.d.). https://www.ibm.com/topics/knn

[5] Band, A. (2022, April 18). *How to find the optimal value of K in KNN? - Towards Data Science*. Medium. https://towardsdatascience.com/how-to-find-the-optimal-value-of-k-in-knn-35d936e554eb

[6] *How SVM Works*. (n.d.). IBM. https://www.ibm.com/docs/en/spss-modeler/saas?topic=models-how-svm-works

[7] *SVM Hyperparameter Tuning using GridSearchCV*. (2020, March 10). Velocity Business Solutions Limited. https://www.vebuso.com/2020/03/svm-hyperparameter-tuning-using-gridsearchcv/