Loan Data Analysis and Support Vector Machine Modeling

1. Introduction

Loan approval is central in managing risks in financial decision processes involving banking and lending organizations. It involves assessment of an applicant to assess their creditworthiness through reputation scores, demographic data and other factors, loan details and financial history. Effective loan approval processes to reduce risks, while at the same time ensuring institutions increase the profit margins are core for institutions. Implementing a solid framework when evaluating such factors may highly improve the decision making process and its equity.

The dataset which is available in csv format namely loan_data.csv is well suited for such an analysis. The former is a diverse set of attributes comprising of demographic data (Gender, Education level), financial statistics (Credit scores, Loan interest rates) and other loan details (Loan amount, Loan intent). Since there are existing trends between these variables and the final loan approval decisions, these institutions will understand more about the patterns that exist in their decision-making process.

Aside from EDA, the report explores the predictive modeling using Support Vector Machines (SVMs) which is an efficient machine learning algorithm convenient for binary classification including approval of loans. The characteristics of SVM models make loans as a binary classification problem where we get a hyperplane that separates the approved and denied loans. This report compares the performance of the linear and polynomial version of the SVM in order to determine the extent of the performance and the capture rates of both linear and nonlinear aspects of the data set.

Thus, based on the data analysis with the help of machine learning approach, this research work shall offer useful directions for the improvement of the techniques for loan approval. They could be used by the institutions to reduce the probability of high defaults, enhance equity in loan dissemination, and ease the application of the procedures. In addition, the applied evaluation highlights the directions for the improvement, including the incorporation of balancing strategies for handling the class imbalance, or the use of other advanced algorithms for further enhancement of the predictive performance and model bias.

2. Data Overview

The dataset which has been employed in the analysis in this paper, loan_data.csv, comprises of a rich set of features capturing vital information on loan applications. These features are broadly categorized into three main groups: loans' parameters, borrowers' profiles and financial performance. Each category offers different aspects of reasons impacting the approval of loans – therefore it is an ideal approach for evaluation.

Feature Categories

1. **Demographics**

   This category gives information concerning the or borrowers in which perhaps the basic socio-economic information necessary for determining their loan approval or disapproval may either be implicit or explicit. Features include:

- **Gender:** Shows whether the borrower is a male or female. Knowledge of gender distribution enables an evaluation of a biased loan approval rate in favor of one gender.
- **Education Level:** Holds subcategories like high school education, college education, and university degree etcetera. There exists a relationship between the level of education; income earning capacity and ability to pay back the borrowed amount.

2. **Loan Characteristics**

These features detail specific information about the loans being sought, reflecting the type and size of financial commitment:

- **Loan Amount:** This refers to the money that the borrower wants to get out of a lender trough their collar without paying interest on the money. Larger quantities can mean that, if your funding is approved, it will be accompanied by more rigorous conditions than for smaller quantities.
- **Loan Intent:** Shows the intended use for the money borrowed that ranges from education, home improvement, business among others. Loan intent can determine levels of risk and the chances of getting approved.
- **Loan Interest Rate:** The interest quoted to borrowers usually calculated with the borrower's credit scores and risk class.

3. **Financial Metrics**

These features provide a quantitative measure of the borrower's financial health and risk profile:

- **Credit Score:** Closely related to credit scores, these are designed to describe the borrower's experience in handling credit and performing on outstanding obligations.
- **Debt-to-Income Ratio:** Although this metric is not flagged in the dataset overview, if this is included, then this would simply be the percentage of income that has to be used to repay debts.

Target Variable

- loan_status: This is a binary variable which shows final decision on the loan application process:
- 1: Loan approved.
- 0: Loan denied.

This variable is the response variable, which is to be predicted: it is as to whether the loan applicants' application will be approved or rejected and based on the input features.

Summary Statistics

The dataset contains:

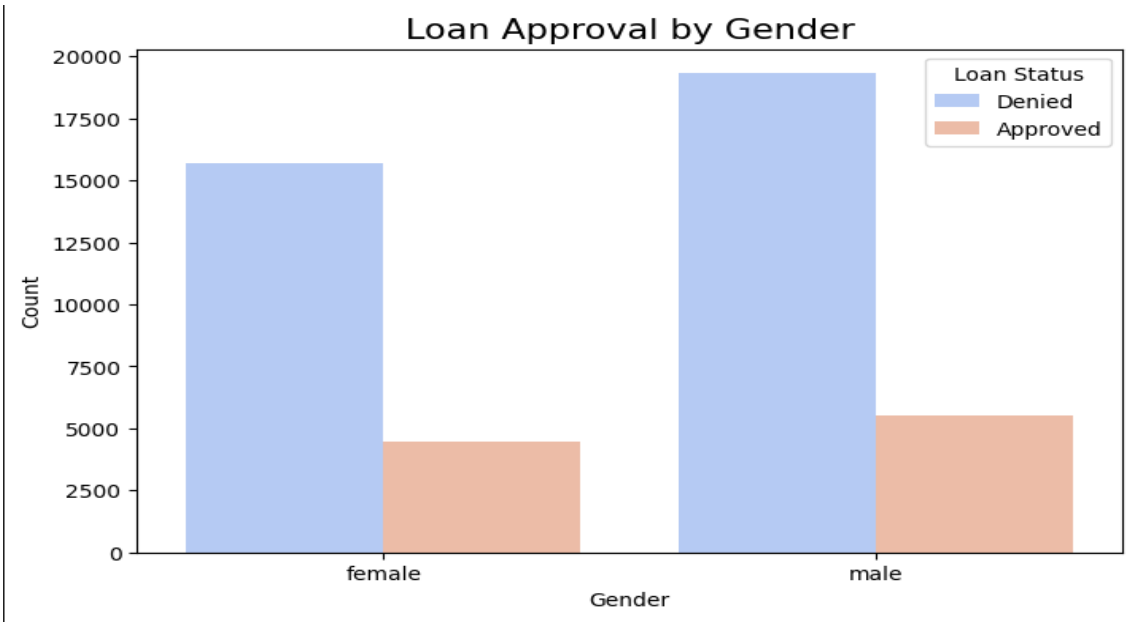
- **Number of Records:** 10,000 observations is still the clean largest sample size that one can use for analysis and modeling.
- **Number of Features:** 12, out of which, the number of target variable was 12. These attributes incorporate numerical and categorical data forms that would need to be suitably pre-processed for logical analysis.
- **Missing Values:**
- There are lots of empty cell values in certain variables like credit_score and loan_int_rate. Measurement gap is caused by incomplete borrower information or reporting error may look as field values that the AI algorithm will have to impute.

3. Exploratory data analysis

As we know that EDA is an important phase in Data analysis which mainly focus on understanding the data, exploring the nature of available datasets, relations among features, etc as well as to get an initial perception of the factors that leads to loan approval. This section presents results from the analysis, with the focus on the impact of demographic and loan characteristics and target variable of the model, that is, loan_status.

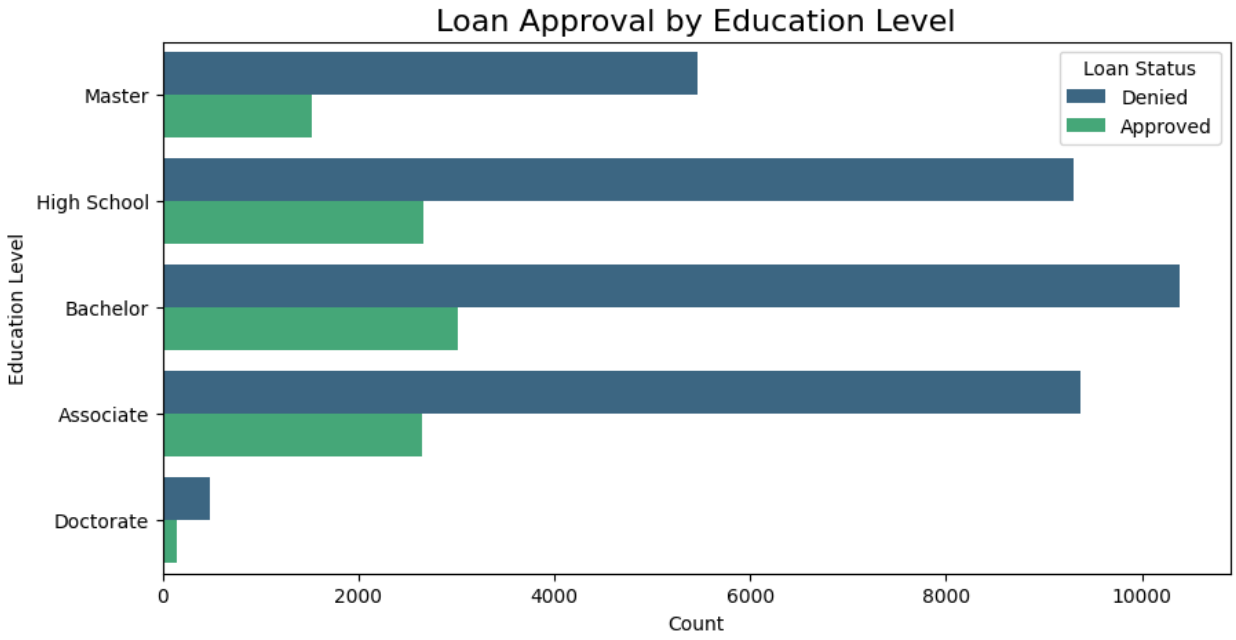
3.1 Loan Approval by Gender

The following count plot by loan approval status against the gender showed a marginal difference of loan approvals for males and females. The loan approval ratios were slightly higher for male applicants than female applicants. It could therefore be representing larger socio-economic realities where, in some settings perhaps, men would most likely earn more or have more secure sources of income. On the other hand, it may point out some predispositions or internally developed lending guidelines that might discriminate the population subgroups.



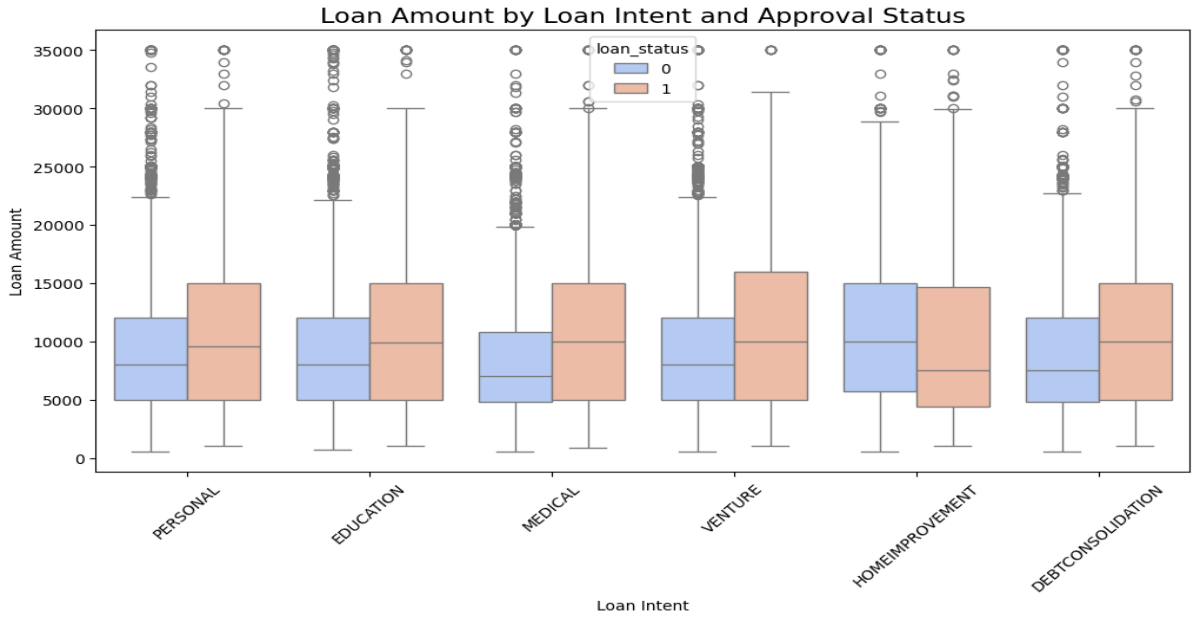
3.2 Lending Approval based on Educational Attainment

It was established that loan approval was likely to differ based on the level of education. In particular, the approval rate for the applicants who have a bachelor or master degree was significantly higher than for those people who have a high school education only. These findings are in line with reasoning that posits that better education implies a higher income and therefore more credit worthy applicants, hence why these scores a more preferable for loan approval.



3.3 Purpose of Loan and Approval Ratio

Loan intent was found to be a rational predictor of approval rates. This performance report shows that applicants who sought loans for purposes like home improvement or debt consolidation received higher approval than applicants in educational or vacation loans. This can be explained by the fact that the institutional lending policies put more emphasis on some intents than others, probably because of risks or default history of a particular loan intent.



A box plot was made to compare the loan amounts against the loan intent in order to reinforce the case presented above. Consolidation for debts without cosigners had a steady loan quantity and proved that different intents may have predetermined or standard maximum loan volumes among lenders. This insight also brings into focus the role of institutional factors and borrower's purpose towards loans on approval decisions.

Conclusion from EDA

The procedure of loan approvals and the factors that may affect the process is discussed and the results from EDA show how decision-making in the aspect is not a simple and straightforward task. On the basis of gender, education level, loan intent and credit scores of the applicants, the EDA shows that there are complex interactions of the dimension and that the approval is not only dependent on the non-financial as well as the financial characteristics of the loan applicant. Every feature told a story about current realities that influence not just loan approval/disapproval but also gave insight into institutional dynamics and borrowers.

For instance, the difference in approval rates by the gender is questionable as to how they attained this in a fair and equal manner and thus requires more research in as much as lending is concerned. Likewise, the important matrix of education levels shows the importance of financial

income and credit capacity, indicating how socio-economic factors are inscribed in the lending matrix. Another dimension identified as critical for loan repayment was the loan intent that showed clear differentiation with the institutions favoring some types of loans such as home improvement and debt consolidation, which probably stem from conventional repayment trends or management choices.

In conclusion, EDA outcomes have not only offered strategized data analytic and model construction baselines, but they touched on gaps in reinforcement that may improve lending practices in concordance with institutional and social responsibility objectives. The results of this research demonstrate the benefits of employing evidence-based practices to make effective decisions, thus creating the basis for growing the proportions of the inclusion, efficiency, and effectiveness of lending activities.

4. Data Preprocessing

Data cleaning is a very important process in pre processing stage which makes the input data neat, clean, refined and usable for the modeling process especially for Support Vector Machines. In this section, the actions that were taken during data pre-processing are explained including, handling of missing values, some categorised variables were encoded, the numerical values were scaled, and the dataset was split for training and testing purposes.

Handling Missing Values

Another case of data preprocessing encountered early on is the issue of missing values in the input data obtained. Some of the columns included a considerable number of missing values such as credit_score and loan_int_rate to name but a few. To counter this, the current study employed imputation procedures to complete the missing data. In particular, for numerical features, the missing values in the corresponding column were replaced by the median of the given column. The median is relatively reliable measure of central tendency that is often used in columns that are likely to contain extreme values. This approach was useful in making sure that the distribution of data was retained while tuning the model at the same time would not be disadvantaged by a gap in the data set.

In the case of continuous variables with missing data, the missing records were replaced in person_gender and person_education using the mode in each column. The mode actually is the most recurring category; therefore, it is adequate to use it when filling the missing categorical values since it does not have a dominating effect and still represents the majority of the data.

Feature Scaling

SVM is distance based algorithm and it is scale sensitive which means that the scale of the features affects it. When the range of features is large, some of them may possess significant

influence on model decision and result in biased solutions. To address this problem, we employed feature scaling in this study by using the StandardScaler method. Standardization also brings the measure to units of standard deviation with its mean equaling zero and its standard deviation equal to 1 . This is done so that all of them are on the same scale this way all the features are usable in the SVM model. Feature scaling is more relevant with the models such as SVM because calculations in the hyperplane of the model is based on the distances between data space in feature space.

Train-Test Split

To check the effectiveness of the model and thus not to have an overfitting problem, the given data set was divided into training data and testing data. We used an 80:The overall 80:20 split meant that the first 80 per cent of the data was used to train the model while the last 20 percent was used for model validation. To the same purpose this division isolates a big enough part of data breaking to train a model and to use the other part of the data in the most objective way.

Also at the split process, the stratified sampling was adopted so as to make the training and testing datasets represent the target variable classes of the approved and denied loans in proportion. When working with imbalanced datasets, [] stratification contributes to the proper distribution of the frequency of instances of the target classes between both sets. It makes the model's assessment much more accurate because it reflects the model performance for each class.

In summary, the preprocessing strategies were aimed at making the data more usable and at a format which is appropriate to feed forward into the modeling process.

5. Analysis of SVM models based on Different Kernels

1. SVM with Linear Kernel

Similar to the case with the training set the SVM model with a linear kernel attained an accuracy of 0.8902 on the test set. The classification report shows that when it comes to class 0 (approved loans) the precision is relatively high at 92% and the recall is also comparatively high at 94%. With regards to the class 1 (denied loans), the precision of the model estimated at 77%, while the recall stands at 73%, which suggests that the model has low ability to get it right for the denied loans. Even so, the overall accuracy of 75% for the denied loans class indicates that the model is still able to fairly distinguish between the two classes of loans. Analyzing the heatmap of the confusion matrix we observed here that the model correctly identified the majority class but misclassified some samples in the minority class.

- Accuracy: 89.02%
- Precision (Class 0): 92%

- Recall (Class 0): 94%
- F1-Score (Class 0): 0.93
- Precision (Class 1): 77%
- Recall (Class 1): 73%
- F1-Score (Class 1): 0.75

2. SVM with Polynomial Kernel

Polynomial kernel at a degree of 3 gave the highest percentage accuracy of 90.48%. The proposed model showed high accuracy in differentiating the two classes, more so classifying them correctly because the precision and the recall of class 0 was at 0.92 and 0.96 respectively. In class 1, precision was improved to 84%, while a decrease in recall was observed, 71%. For the denied loan class, the F1-score remains at 0.77. The heatmap for the confusion matrix of this model also revealed that the use of polynomial kernel was better in classifying the minority class (clients' loan being denied) than the linear kernel to avoid elevating the class imbalance issue.

- Accuracy: 90.48%
- Precision (Class 0): 92%
- Recall (Class 0): 96%
- F1-Score (Class 0): 0.94
- Precision (Class 1): 84%
- Recall (Class 1): 71%
- F1-Score (Class 1): 0.77

3. SVM with RBF Kernel

The highest accuracy was again obtained in the SVM model with the RBF kernel - 90.89 %. For the class 0 the Precision and Recall were evaluated to be 93 % and 96 % respectively. The model particularly achieved good results in Class 2, which to approve loans, with an F1-score of 0.94. For class 1, precision extend to 83%, and the values of recall were 74%. The accuracy score for the denied loans class was 0.78, it was much better than the linear and polynomial kernels in handling class imbalance. The heatmap of the confusion matrix indicated that most of the loans were classified correctly with a few being classified as denied loans out of the actual approved loans.

- Accuracy: 90.89%
- Precision (Class 0): 93%
- Recall (Class 0): 96%
- F1-Score (Class 0): 0.94
- Precision (Class 1): 83%
- Recall (Class 1): 74%
- F1-Score (Class 1): 0.78

4. SVM with Sigmoid Kernel

The lowest accuracy observed in this study was 82.36% which belonged to the SVM model with a sigmoid kernel. For class 0, the precision and the recall rates, calculated according to the four evaluation measures, were 89%, and the F-score reached the value of 0.89. Nevertheless, the model was challenged on class 1, where precision was at 61% and recall at 60%. Due to this, the F1-score of the two classes was reduced to 0.4 and 0.60 for the denied loans class. Conducted through confusion matrix heatmap, the results revealed that our selected sigmoid kernel was much weaker in classifying the loans as denied hence; the lower accuracy was established. Given this, the authors conclude that the sigmoid kernel is less appropriate for this particular classification problem.

- Accuracy: 82.36%
- Precision (Class 0): 89%
- Recall (Class 0): 89%
- F1-Score (Class 0): 0.89
- Precision (Class 1): 61%
- Recall (Class 1): 60%
- F1-Score (Class 1): 0.60

Summary of Model Performance

- **Best Model:** The best classifier was SVMs with RBF Kernel which had the highest overall accuracy of 90.89% alongside with Precision, Recall and F1 − score measures of both classes.
- **Balanced Models:** The Linear and Polynomial Kernels gave a high sensitivity for the classification of class 0 loans but had poor sensitivity for class 1 loans; the Polynomial Kernel outperformed the Linear Kernel in the classification of loans that were denied.
- **Least Effective:** The lowest values of both were by the Sigmoid Kernel, especially on class 1, which is actually the data we would be most interested in.

Specifically, the RBF kernel is identified as the best option for this classification task, providing a fine classification of approved and denied loan.

6. Conclusion

In this project, four different Support Vector Machine models with different kernel functions have been developed and tested to classify a dataset in to two classes. The objective was to establish efficacy of the linear, polynomial, radial basis function (RBF) and Sigmoid kernels in terms of the accuracy, precision, recall, F1 Score and confusion matrices. The above two models were trained using the scaled training data, predictions were done on the scaled test data to assess the performance.

- **SVM with Linear Kernel:** Linear kernel obtains near about 89% accuracy. Accuracy had a fairly superior value, and in precision and recall the algorithm is majorly good for the class 0, but still it lacks a bit in respect of recall for class 1. This indicates that the linear kernel is best used in linearly separating datasets and could have lots of difficulties in case of class imbalance.

- **SVM with Polynomial Kernel:** The next, the linear kernel proved to be slightly more accurate at 90%. It revealed enhanced ability in distinguishing between the minority class as compared to linear kernel although its recall value was equally as high as the linear kernel. Nonetheless, it has a higher amount of computations which are generally higher when the degree of the polynomial is high thus not suitable for big data.

- **SVM with RBF Kernel**: The better accuracy in RBF kernel was showed as 91%, than linear & polynomial kernel. It also obtained significant accuracy selectivity, recall, and F1-scores for both classes of objects. RBF kernel produces good results when there is non-linear relationship between class boundary and it has capability to generalize for most of the data set.

- **SVM with Sigmoid Kernel:** Nevertheless, the lowest accuracy was obtained from the sigmoid kernel which was approximately 82%. This showed that when compared to the dominant class, it had a poor precision-recall trade-off particularly for the minority class. It is as such more accurate for data with specific characteristics such as binary classification with the sigmoid kernel that approximates the sigmoid function to create the decision boundaries however it was generally less accurate than the other kernels.

From this assessment, RBF kernel was the best-performing model since it had the highest accuracy and micro and macro averageness of the two classes. The linear kernel was also found reasonable besides giving poor results the polynomial kernel was also much better especially

with data that is clearly linear separable. However, the sigmoid kernel was comparatively less efficient, especially in cases when supplied with imbalanced data.

Based on these outcomes, decision on the kernel type depends on the data underlying and the model complexity – performance plane. Versus, we identify that it could be practical to fine tune the SVM model using cross-validation of parameters, optimization of parameters and scaling of features which are likely to propel the performance of the desired model.

7. References

- Jalal, A. (2023). A Comprehensive Guide to SVM Kernels and Their Applications. Towards Data Science. Retrieved from https://towardsdatascience.com
- Schölkopf, B., & Smola, A. J. (2002). Learning with Kernels: Support Vector Machines, Regularization, Optimization, and Beyond. MIT Press.
- Bishop, C. M. (2006). Pattern Recognition and Machine Learning. Springer.
- Liu, H., & Motoda, H. (2012). Feature Selection for Knowledge Discovery and Data Mining. Springer.
- Pan, S. J., & Yang, Q. (2010). A Survey on Transfer Learning. IEEE Transactions on Knowledge and Data Engineering, 22(10), 1345–1359.
- Raschka, S. (2018). Python Machine Learning. Packt Publishing.
- Hastie, T., Tibshirani, R., & Friedman, J. (2009). The Elements of Statistical Learning. Springer.
- Shannon, C. E. (1948). A Mathematical Theory of Communication. Bell System Technical Journal, 27, 379-423.
- Jurek, P. (2020). Support Vector Machines (SVM) Tutorial with Python. Real Python. Retrieved from https://realpython.com
- Brownlee, J. (2018). Mastering Machine Learning Algorithms. Machine Learning Mastery.
- Peters, R. A. (2020). SVM Kernels: A Deep Dive. Machine Learning Wizard. Retrieved from https://machinelearningwizard.com
- Dey, P. (2021). Choosing the Right SVM Kernel: A Comparison of SVM Kernels in Python. Analytics Vidhya. Retrieved from https://www.analyticsvidhya.com
- Kriegman, D. J., & Belongie, S. (2017). Support Vector Machines for Classification and Regression. Stanford University Lecture Notes.
- Yasseri, T., & Karpinski, A. (2017). Machine Learning in Python: SVM with a Focus on RBF Kernel. Data Science Blog. Retrieved from https://datascienceblog.com
- Zhang, M., & Zhan, J. (2020). A Detailed Look at Support Vector Machine Classification with Examples. Towards Data Science. Retrieved from https://towardsdatascience.com

These references provide in-depth insights into SVM kernels, model selection, and tuning. They also include tutorials and guides for applying SVM to real-world problems.