SVM-Based Prediction of Above-Median Wine Purchasers

Dikun (Daisy) Fang
Dyson School of Design Engineering, Imperial College London · DESE40009 Data Science, 2026

Individual section of the group report "Predicting Customer Purchasing Behaviour" (H. Bhudia, S. Talekar, S. Xie, D. Fang), reproduced as a stand-alone document.

***Abstract*—Our group asked how effectively machine learning methods can identify actionable retail customer patterns, with each member applying a different method to a different question in the Customer Personality Analysis dataset [1]. This section uses a Support Vector Machine (SVM) to identify customers whose wine expenditure is above the sample median from demographic and purchasing-channel records. A tuned RBF SVM (C = 0.5, gamma = 0.02) achieved test accuracy 0.938, precision 0.922, recall 0.955, F1 0.939 and ROC-AUC 0.980, the strongest result among the four methods in the group study.**

I. Introduction

This investigation addresses whether demographic and purchasing-channel records can identify existing customers whose wine expenditure is above the sample median, supporting more selective wine promotions. Customer purchasing behaviour varies across product categories, so a uniform marketing strategy may not be equally effective for every customer. Sending the same wine promotion to all customers can waste resources and produce limited returns. Identifying customers with stronger wine-purchasing behaviour can therefore support customer segmentation, personalised communication and more efficient allocation of marketing resources [2].

II. Methodology

Support Vector Machine (SVM) is a supervised learning method for classification and regression. In binary classification, it identifies a hyperplane that separates two classes while maximising the margin between the nearest observations, known as support vectors [3]. A wider margin is intended to improve generalisation, while permitting some classification errors when the classes are not perfectly separable. Kernel functions allow SVM to represent nonlinear boundaries without explicitly constructing interactions [4]. This is relevant because wine expenditure may reflect nonlinear relationships among income, household structure and purchasing-channel activity. The hyperparameters are:

- **C:** controls the trade-off between margin width and classification error. Larger values penalise errors more strongly and can increase overfitting.
- **Gamma:** controls the radius of influence of each training observation for the RBF kernel. Large values permit highly local and potentially overfitted boundaries.
- **Kernel:** determines the transformation used to construct the decision boundary; linear, RBF, polynomial and sigmoid alternatives were compared.

SVM is appropriate because the outcome is binary, the predictors contain multiple demographic and behavioural dimensions, and their relationship with wine expenditure need not be linear. The limited number of hyperparameters also permits sensitivity analysis. F1 was prioritised because precision represents targeting efficiency, while recall represents coverage of above-median purchasers. Hyperparameters were selected using validation data only, and the test set was reserved for final evaluation [4].

III. Data Preprocessing

The Customer Personality Analysis dataset contains 2,240 customers and 29 variables [1]. Following the shared group pipeline, three records with Year_Birth before 1930 and one Income value of 666,666 were removed. Twenty-four missing Income values were imputed with the training-set median. Marital labels "YOLO," "Absurd" and "Alone" were consolidated into "Single," while constant columns Z_CostContact and Z_Revenue were excluded. The raw audit identified 182 repeated profiles after excluding ID; none were removed because the group applied no deduplication rule. Label consolidation produced 184 repeated cleaned profiles. The final sample contained 2,236 customers.

Age at enrolment was calculated from the enrolment year and Year_Birth. The target was derived from MntWines, representing expenditure during the previous two years. The cleaned median was 174: values above 174 were labelled 1 (1,117 customers), and values at or below it were labelled 0 (1,119). This defines an above-median segment rather than a profit-based "high-value customer."

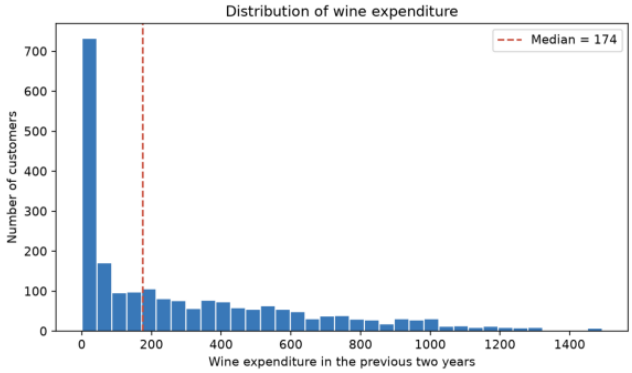


Fig. 1. Distribution of wine expenditure and the cleaned-sample median threshold.

Twelve original predictors were retained: Income, age at enrolment, Education, Marital_Status, Kidhome, Teenhome, Recency, NumWebVisitsMonth, NumDealsPurchases, NumWebPurchases, NumCatalogPurchases and NumStorePurchases. These represent purchasing power, lifecycle and household structure, recency, digital engagement and purchasing frequency across channels. MntWines was

excluded because it defines the target; other product-expenditure variables were withheld so the analysis tested demographic and channel information rather than closely related spending totals. Education and Marital_Status were one-hot encoded, producing 20 transformed features. Numeric variables were standardised because distance-based margins are scale-sensitive [4]. Imputation, encoding and scaling were fitted inside a scikit-learn training pipeline to prevent leakage [5]. Stratified splitting produced 1,341 training, 447 validation and 448 test observations.

IV. Hyperparameter Selection and Sensitivity Analysis

Linear, RBF, polynomial and sigmoid kernels were compared using C = 1 and gamma = scale. RBF achieved the highest SVM validation F1 (0.930), compared with polynomial (0.923), linear (0.919) and sigmoid (0.907; Fig. 2). RBF was therefore selected because it provided the strongest validation balance within the SVM family while allowing a smooth nonlinear boundary.

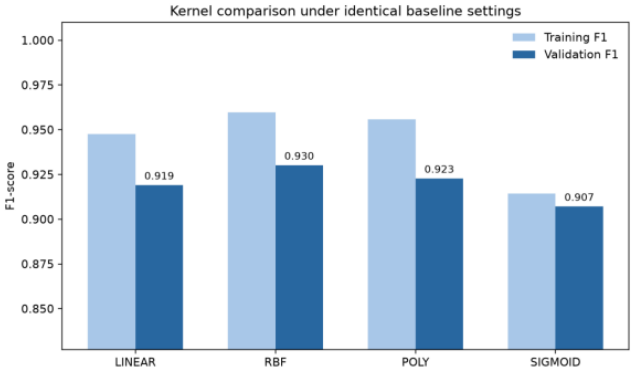


Fig. 2. Training and validation F1 for alternative SVM kernels.

With RBF fixed, C values from 0.05 to 20 were evaluated. Validation F1 was numerically highest at C = 0.5 (0.935), while training F1 increased from 0.930 to 0.996 as C increased (Fig. 3). The absence of validation improvement at larger values, together with the widening training-validation gap, indicated increasing overfit. C = 0.5 was retained. Gamma was then varied from 0.001 to 1. Gamma = 0.02 achieved the highest validation F1 (0.935), narrowly exceeding gamma = 0.1 (0.934), but with a smaller training-validation gap. At gamma = 1, training F1 reached 0.993 while validation F1 fell to 0.831 (Fig. 4). The locked configuration was therefore RBF, C = 0.5 and gamma = 0.02.

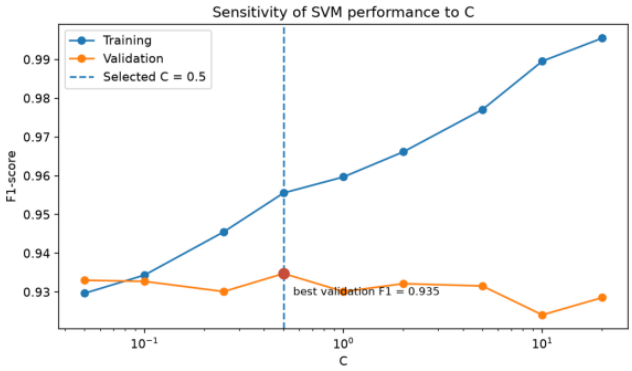


Fig. 3. Training and validation F1 across C values.

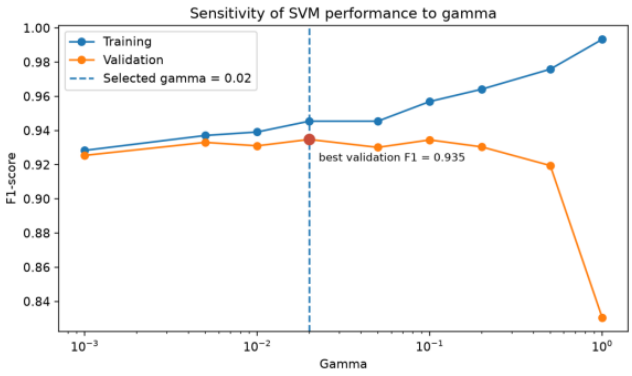


Fig. 4. Training and validation F1 across gamma values.

V. Results

The final SVM showed consistent performance across splits (Table I). Test accuracy was 0.938, precision 0.922, recall 0.955, F1 0.939 and ROC-AUC 0.980. ROC-AUC measures ranking performance across possible thresholds [6], whereas F1 evaluates the locked binary decisions; reporting both avoids relying on accuracy alone. The confusion matrix contained 206 true negatives, 18 false positives, 10 false negatives and 214 true positives (Fig. 5). Thus, 95.5% of the actual above-median purchasers were identified, while 92.2% of customers predicted as above-median belonged to that segment.

Table I. Final SVM Performance

Subset	Accuracy	Precision	Recall	F1	ROC-AUC
Training	.944	.922	.970	.945	.985
Validation	.933	.907	.964	.935	.973
Test	.938	.922	.955	.939	.980

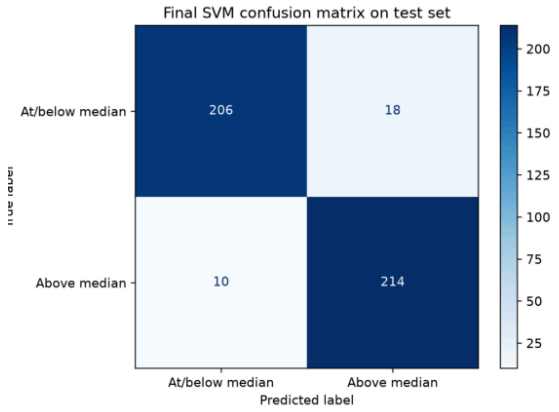


Fig. 5. Confusion matrix for the final SVM on the held-out test set.

VI. Discussion

For targeted marketing, the 18 false positives represent potentially unnecessary contacts, whereas the 10 false negatives represent missed above-median purchasers. The appropriate operating threshold should therefore depend on campaign cost and the expected value of a successful contact. Permutation importance, computed by shuffling one original predictor and measuring the reduction in test F1 [7], ranked NumWebPurchases, NumCatalogPurchases, NumStorePurchases and Income highest (Fig. 6). The result indicates that cross-channel purchasing intensity was more informative to the fitted model than most demographic variables. It describes predictive dependence, not causation.

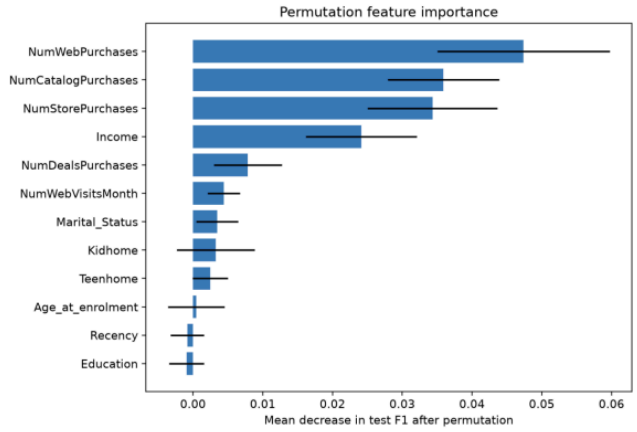


Fig. 6. Permutation importance measured by the decrease in test F1.

The analysis has four limitations. First, a median threshold is statistically convenient but does not incorporate margin, campaign cost or customer lifetime value. Second, channel counts and MntWines cover the same historical period, so the model performs segmentation rather than future forecasting. Third, repeated profiles were retained for consistency with the group pipeline and may increase the influence of common profiles. Fourth, one split does not quantify all sampling uncertainty. Future work should use time-ordered outcomes, external validation, repeated cross-validation, probability calibration and cost-based threshold selection. Within these limits, the tuned RBF SVM provides a reproducible method for identifying existing above-median wine purchasers in this cleaned customer sample.

References

[1] I. Akash, "Customer Personality Analysis." [Online]. Available: https://www.kaggle.com/datasets/imakash3011/customer-personality-analysis

[2] M. Wedel and W. A. Kamakura, *Market Segmentation: Conceptual and Methodological Foundations*, 2nd ed. New York, NY, USA: Springer, 2000, doi: 10.1007/978-1-4615-4651-1.

[3] C. Cortes and V. Vapnik, "Support-vector networks," *Machine Learning*, vol. 20, pp. 273–297, 1995, doi: 10.1007/BF00994018.

[4] T. Hastie, R. Tibshirani, and J. Friedman, *The Elements of Statistical Learning*, 2nd ed. New York, NY, USA: Springer, 2009, doi: 10.1007/978-0-387-84858-7.

[5] F. Pedregosa et al., "Scikit-learn: Machine Learning in Python," *Journal of Machine Learning Research*, vol. 12, pp. 2825–2830, 2011.

[6] T. Fawcett, "An introduction to ROC analysis," *Pattern Recognition Letters*, vol. 27, no. 8, pp. 861–874, 2006, doi: 10.1016/j.patrec.2005.10.010.

[7] A. Fisher, C. Rudin, and F. Dominici, "All models are wrong, but many are useful: Learning a variable's importance by studying an entire class of prediction models simultaneously," *Journal of Machine Learning Research*, vol. 20, no. 177, pp. 1–81, 2019.

Code

github.com/dfangniv/DE1-DESE4009-Data-Science

Declaration of Generative AI Use

Generative AI tools were used to support language editing, improve clarity, and condense the report to meet the required word limit. They also assisted with code troubleshooting and document formatting. The research questions, analytical approach, feature selection, modelling decisions, interpretation of results, and conclusions were developed by the group. All AI-assisted content and code were critically reviewed, verified against the original data and model outputs, and approved by the group members.