Comparative Analysis of Machine Learning Methods for E-Commerce Analytics: Sentiment Classification and Customer Churn Prediction

Saatvik Reddy Gutha
Student ID: x24257460
x24257460@student.ncirl.ie

Dattathreya Chintalapudi
Student ID: x24212881
x24212881@student.ncirl.ie

***Abstract*—The critical motivations of the research were to find out how machine learning can be applied to E-Commerce analysis on two types of data, namely, text, which is taken from product reviews, and structured customer behavioral data. The algorithms that were tried in this particular work include Naive Bayes, Support Vector Machines, Random Forest, and Extreme Gradient Boosting. The project followed the CRISP-DM data mining process. We considered a dataset of more than 568,000 Amazon Fine Food Reviews for sentiment classification and 10,000 Bank Customer Records for churn prediction. The TF-IDF vectorization technique was used for converting the reviews to classify sentiments into either positive, neutral, or negative. Predicting customer churn would include finding out demographics and behaviors in the data that determine whether a customer is at risk.**

Results showed that SVM outperformed Naive Bayes in text classification, with an accuracy of 78.9% and an F1-score of 0.782. In predicting churn, XGBoost had a slight edge over Random Forest, with an accuracy of 86.7% and 0.869 ROC-AUC. Using SHAP and LIME interpretation tools provided valuable business insights. They illustrated the impact of certain word patterns on review sentiment and the fact that customer churn is predominantly driven by age, number of products purchased, and geographical location.

***Index Terms*—Machine Learning, E-Commerce Analytics, Sentiment Analysis, Customer Churn Prediction, CRISP-DM, Text Classification, SHAP, LIME, Random Forest, XGBoost**

I. Introduction

Anxiety and depression, traditionally viewed as disorders of a single individual, can also occur within a social network.

The rapid transformation of retail through e-commerce has generated volumes of data on customers, including reviews, purchase history, and behavioral habits, that have never been seen before. By applying machine learning to parse this material, firms are able to track customer sentiment, anticipate churn, and tailor retention strategies (Chapman, 2000). This research addresses two key analytics problems in e-commerce that have significant business consequences.

Basic challenges involve sentiment analysis of product reviews to automatically provide insight into customer satisfaction, product quality issues, and brand perception. Due to the many thousands of reviews generated each day across the majors, manual observation is out of the question, and so machine learning is essential for any scalable sentiment monitoring. Another challenge involves predicting customer churn in advance of actual defection to allow targeted retention campaigns; doing this reduces customer acquisition costs, typically five to twenty-five times that of retention. [9].

A. Research Objectives

The primary objectives of this study are:

1) It aims to critically evaluate machine learning methods across text-based product reviews and structured customer behavior data.
2) Apply and compare at least two different algorithms on each dataset using the CRISP-DM methodology.
3) To implement the model interpretability techniques, SHAP, and LIME, that provide actionable business insights.
4) Identifying important predictors of customer sentiment and churn behavior.
5) Provide evidence-based recommendations for e-commerce analytics implementation.

B. Research Questions

This study addresses the following research questions:

- RQ1: How do traditional classifiers (Naive Bayes, SVM) compare for e-commerce sentiment classification?
- RQ2: Which ensemble method (Random Forest, XG-Boost) performs better for customer churn prediction?
- RQ3: What features are most predictive of customer churn?
- RQ4: How can model interpretability techniques enhance business decision-making?

This paper is organized as follows: The second section reviews the literature related to sentiment analysis and churn predictions, with a critical evaluation of the usefulness of this literature. The third section gives a detailed account of the components of our as described by the CRISP-DM Methodology and the Data Preparation phase. The fourth section includes a discussion of evaluation metrics, sampling techniques and results of our study. The fifth section presents our findings and discusses their implications for use within a business context. Finally, the sixth section will summarize this study by discussing its limitations and future research recommendations.

II. Related Work

This section looks at previous research in the fields of: sentiment analysis; churn prediction; and model interpretability. It examines how well each publication solves our project's problem and identifies any gaps in knowledge that we intend to fill.

A. *Sentiment Analysis in E-Commerce*

Pang and Lee [2] provide a seminal review of opinion mining and sentiment analysis, showing that for any task where labeled data exist, supervised machine learning methods outperform lexicon-based methods. Their work is helpful in defining the theoretical boundaries of sentiment classification, but their research is prior to state-of-the-art e-commerce review datasets and does not directly address the specific issues of product review sentiment, such as domain-specific vocabulary and complex merged sentiment expressions. The current study expands on their work in overcoming the deficiencies by adopting TF-IDF with bigrams to grasp negation patterns.

The authors Liu [3] introduced Aspect-Based Sentiment Analysis (ABSA), which is useful for understanding how people encode different opinions in one review. One of the key findings from Liu's work is that multiple nuances of sentiment exist within an individual review. For instance, in regards to a product, many customers will have expressed an overwhelming positive opinion towards the product's quality, even if they have expressed an overwhelmingly negative opinion towards the product's price. While Liu's research primarily emphasises the theoretical aspect of ABSA, the focus of our research is on applying these concepts at a large-scale through the application of computationally effective classic classifiers.

As cited in McCallum (1998), the researchers conducted a study comparing the Event Models used in the Naive Bayes approach to text classification. Their conclusion was that the Multinomial Model consistently performed significantly better than the Bernoulli model when applied to longer documents. Thus, their conclusion helped to determine which Model we should use for Amazon reviews, which average 40+ words - Multinomial Naive Bayes. One limitation of McCallum and Nigam's study was that it focused primarily on the classification of news articles, while informal reviews will have very different linguistic structures. However, McCallum and Nigam's study serves as an excellent benchmark for performance when comparing different Event Models when classifying high-dimensional sparse text.ance for high-dimensional sparse text data.

Cortes and Vapnik [5] introduced Support Vector Machines, demonstrating effectiveness for high-dimensional classification. Joachims [6] extended SVMs specifically to text categorization, showing that linear SVMs achieve competitive performance with computational efficiency. The strength of these works lies in their theoretical rigor and proven performance; however, they predate deep learning advances. For our study, we deliberately chose traditional classifiers over deep learning to prioritize interpretability and computational efficiency for business deployment.

The researchers analyzed the Amazon Fine Food Reviews dataset, which also served as the basis for our study, to understand how reviewers acquire expertise over time based on the number of reviews they have posted about a certain product. By identifying the distribution by stars across time, this research confirms our findings that show there is a class imbalance present in people's ratings on Amazon (64% were rated positively). Although this research does provide some insight into rating distributions, it does not directly address the performance of methods used to classify the actual sentiment behind these reviews; however, this is something that we aim to address in our study.

In their paper, Zhang et al. [3] compare traditional and deep learning approaches for sentiment classification, with deep learning showing advantages when using large amounts of data, but traditional methods being more interpretable. As such, the balance between interpretability and accuracy has been a major consideration in our approach to using LIME explanatory models to create explanations for our business stakeholders.

B. *Customer Churn Prediction*

Reichheld citereichheld1996loyalty has made an analysis of how reducing churn by just 5

Neslin et al. [10] conducted comprehensive evaluations of the accuracy of customer churn prediction models across many different business sectors, including conducting a meta-analysis to identify the most important considerations when evaluating model performance, such as effective feature engineering and how to effectively handle class imbalance. They provided a framework for evaluating model performance that has influenced our research by using multiple metrics to evaluate performance: precision, recall, F1 score, and AUC. However, as their work was completed before advances in gradient boosting occurred, our study employed the use of XGBoost, which has consistently been shown to provide significantly superior performance in the many competitions to which it has been entered over the past year.

Breiman [11] introduced the concept of Random Forests. Random Forests incorporate the additional benefit of protecting against model overfitting through the implicit quantification of the importance of a given feature through the use of Gini impurity. Random Forests allow for the modelling of non-linear relationships between behavioural variables in the prediction of churn. A disadvantage of Random Forests is that they are sensitive to the selection of hyperparameters (the number of trees, the maximum depth of the tree), which we address through hyperparameter tuning via cross-validation.

Chen and Guestrin [12] developed XGBoost, which incorporates regularisation and enables the efficient implementation of state-of-the-art predictive models. XGBoost has developed a new approach to overcome the limitations of Random Forests, specifically by developing a sequential boosting method that focuses on the most difficult examples. We will be using both Random Forests and XGBoost to compare bagging (Random

Forests) versus boosting (XGBoost) methods for predicting customer churn.

Chawla et al. [13] introduced SMOTE to address class imbalance prevalent in churn datasets where churners represent minorities. SMOTE generates synthetic examples through interpolation, improving classifier sensitivity. A potential limitation is generating unrealistic synthetic examples; we mitigate this by preserving original test distributions for realistic evaluation.

Verbeke et al. [14] compared machine learning algorithms for churn prediction using profit-driven evaluation, finding ensemble methods consistently superior. Their business-oriented evaluation framework aligns with our interpretability focus. However, their telecommunications context differs from e-commerce; our study contributes by applying similar methods to bank customer churn as a proxy for e-commerce customer behavior.

Vafeiadis et al. [15] conducted comprehensive comparison of classification techniques for churn prediction, confirming ensemble method superiority while highlighting computational requirements. This work validates our algorithm selection but does not incorporate SHAP interpretability which we add for business actionability.

C. Previous Uses of Datasets

Sentiment analysis research has relied heavily on the Amazon Fine Food Reviews dataset. Numerous Kaggle competitions have employed various techniques (e.g. Naive Bayes, SVMs and Deep Learning) for sentiment classification. Historical results demonstrate high classification accuracy for binary classifying between 70% - 85% as well as for classifying into three classes between 65% - 75%. Although we have performed a systematic comparison of traditional classifiers to identify the most important features to produce an interpretable prediction, this study also demonstrates that combining classification and interpretability enables us to determine which features contribute most to a classifier's prediction.

The Bank Customer Churn dataset has been extensively benchmarked with Kaggle and reported ROC AUC results ranging between 0.80 to 0.88 using a variety of ensemble-based algorithms. Our research offers the additional contribution of incorporating SHAP analysis to assist us to identify specific customer attributes contributing to customer churn. Our findings enable businesses to better reach targeted customers for future retention efforts, instead of relying solely on the predictions of our models.

D. Model Interpretability

Lundberg and Lee citelundberg2017shap created SHAP, a framework for explaining or interpreting machine learning models by identifying what features contributed most to the predicted outcome of a machine learning model (using Shapley values). The biggest benefit of SHAP is the ability to provide consistent and fairly accurate local explanations, but its computational complexity increases dramatically as dataset size increases. We leverage TreeSHAP, a version of SHAP that was specifically optimized for Tree-Based models, to maintain efficiency.

Ribeiro et al. citeRibeiro2016lime created LIME, which generates local interpretable explanations of individual predictions using perturbations. The main advantage of LIME is that it is model-agnostic, thus being uniquely positioned to be applied to text classification problems by identifying which word(s) influenced the model's predictions. The limitation with respect to LIME is that it may generate unstable explanations with respect to perturbations, and we overcome this limitation by generating and aggregating multiple explanations.

Molnar citemolnar2020interpretable synthesized the state of the art in Explainable AI and provided an in-depth comparison of global and local explanation methods. His work provided us with a framework to identify LIME as the most appropriate method for providing word-level explanations for classification problems, and SHAP for providing feature importance explanations for churn prediction problems.

E. Research Gap

Although the literature review identified advances within sentiment analysis and churn forecasting, ANd therefore lack of comprehensive perspectives most articles entailed an assessment of either sentiment analysis, or churn prediction yet, most studies focus heavily on accuracy, with less of an emphasis on interpretability. This study bridges these gaps by examining both text (sentiment analysis) and structured (churn forecasting) data in a unified analysis using a CRISP-DM based methodology while applying both LIME and SHAP to facilitate obtaining relevant actionable insights for a business through the impact created by sentiment leading to churn.

III. Data Mining Methodology

This study follows the Cross-Industry Standard Process for Data Mining (CRISP-DM) framework [1], encompassing six phases: business understanding, data understanding, data preparation, modeling, evaluation, and deployment.

A. Business Understanding

Two critical business challenges motivate this study:

Sentiment Monitoring: Automated sentiment classification enables stakeholders to respond to negative comments quickly and to track satisfaction levels over time.

Churn Prevention: By identifying customers at risk of churning, businesses can implement follow-up campaigns aimed at retaining at-risk customers. The costs associated with acquiring a new customer are generally between 5× and 25× higher than those associated with retaining a customer. Therefore, the economics of accurately predicting the churn of their customers directly affect businesses' profitability.

B. Data Understanding

1) Amazon Fine Food Reviews (Text): This is a dataset generated by McAuley et al. in 2013 that contains 568454 reviews between 1999 and 2012 with attributes including the review text, summary, star rating (1-5), helpfulness votes and timestamp of review submission. The sentiment of the review can be classified as either positive, negative or neutral based on the rating provided. The class distribution is heavily weighted toward the positive (64%), whereas negative represents 21% and neutral represents 15%.

2) Bank Customer Churn (Structured): In addition, the dataset contains 10000 records containing 14 attributes per customer: Age, Gender, Geography (Geography, Gender, Age), Tenure, Balance, Number of Products, Credit Card, Active Membership, Credit Score and Estimated Salary. It is important to note that customers marked as ”Churn” have a corresponding binary target (1) or ”Retention” (0). Approximately 20 percent of the sample is classified as Churn (Churn).

TABLE I
DATASET CHARACTERISTICS

Attribute	Amazon Reviews	Bank Churn
Instances	568,454	10,000
Features	10	14
Data Type	Text	Numeric/Categorical
Target	3-class sentiment	Binary churn
Class Balance	Imbalanced (64% pos)	Imbalanced (20% churn)

C. Data Preparation

1) Text Preprocessing Pipeline: The authors made six preprocessing steps: (1) Lowercasing all the words in the dataset; (2) Removing all HTML tags and URLs using Regular Expressions (regex); (3) Removing all punctuation/ characters and numbers; (4) Tokenisation using Natural Language ToolKit (NLTK); (5) Removal of stopwords using English stopword lists; and (6) WordNet lemmatisation so that there is a reduced vocabulary.

Parameters for TF-IDF Vectorisation included: (i) maximum number of features = 5, 000 to balance dimensionality and vocabulary coverage; (ii) unigrams and bigrams to capture phrases like ‘not good’; (iii) minimum document frequency = 5 to exclude infrequent or rare words; and (iv) maximum document frequency = 95 per cent to exclude common or high-frequency words.

Sampling Strategy:To preserve the original class distributions and minimize the computational cost, a stratified random sample of 50,000 reviews was created. Stratification permits the retention of proportionate representation of minority classes (neutral and negative).

2) Structured Data Preprocessing: The data cleaning of Bank Churn did not require much effort, as there were no missing values within it. The categorical variables (Geography and Gender) were label encoded. The following features (RowNumber, CustomerId, and Surname) were removed from the dataset, as they did not provide any predictive information for the model. Numeric features were normalised to mean zero and variance of one using StandardScaler, which is important when training models using SVM and helps with the convergence of gradient boosting. **Class Imbalance Handling**: SMOTE generated synthetic minority (churn) examples by interpolating between existing churned customers in feature space. Training data was balanced to 50% churn rate. Original test set distribution was preserved to evaluate real-world performance on imbalanced data.

D. Modeling

Four algorithms were implemented with each member responsible for two distinct models:

Member 1 - Sentiment Analysis: Naive Bayes and SVM

Member 2 - Churn Prediction: Random Forest and XGBoost

Model Specifications and Parameterization:

Multinomial Naive Bayes: Setting Laplace smoothing to alpha=1.0 prevents the observation of zero probabilities for terms that have not been seen in the training data. This choice of parameter balances the risk of either overfitting (using a lower alpha) or underfitting (using a higher alpha). Alpha-1 is the most commonly used value for text classification.

Linear SVM: Setting the regularization value to C=1.0 allows for a balance between finding the largest margin possible while minimizing the cost of misclassifying a sample. Using a smaller C value increases regularization (simpler boundaries) where as using a larger C value allows for more complex boundaries which run the risk of reconciling the two approaches to overfitting. The regularization value of C=1.0 has been found to produce a balanced bias-variance trade-off when verified through cross-validation.

Random Forest: By averaging over 100 random trees, Random Forest produces predictions that are stable and relatively close to what can be expected from other predictors; adding additional trees results in a diminishing return on accuracy. The maximum tree depth is set to 10 so that the depth of the trees does not lead to overfitting, while still allowing the trees to represent non-linear relationships. The deeper the trees, the better they fit to training data; the shallower the trees, the worse they fit to the data.

XGBoost: The learning rate of 0.1 finds a balance between speed of convergence and accuracy of results; for lower learning rates, it is often said that more iterations are required, but these iterations usually produce a better optimum. The maximum depth of the trees produced is 5 which is shallower than the Random Forest trees; boosting introduces additional complexity through iterations and therefore, the shallower trees produce a lower tendency for overfitting. Implementing the L2 regularization value of 1.0 helps in preventing overfitting.

E. Model Interpretability Implementation

LIME for Text: Perturbs input by randomly removing words, observes prediction changes, fits local linear model. Explains which words drive individual sentiment predictions.

SHAP for Churn: TreeSHAP efficiently computes Shapley values for tree ensembles. Provides both global feature importance (mean absolute SHAP) and local explanations (individual customer predictions).

IV. Evaluation

A. Performance Metrics Selection

Multiple metrics were selected to capture different aspects of classification performance:

Accuracy: This will reflect overall correctness but can be misleading in imbalanced datasets where forecasting the majority class yields high accuracy. Included for baseline comparability but not as the primary performance metric.

Precision: Proportion of positive predictions that is correct. Very important in the case of targeting churn, as false positives waste resources in retention campaigns on non-churners.

Recall (Sensitivity): It refers to the share of actual positives which was correctly identified. Very important for identifying potential churners before attrition; missed churners mean lost potential revenue.

F1-Score: The harmonic mean of precision and recall, giving a balanced measure when both false positives and false negatives are costly. Thus, it is the main measure when there is an imbalanced classification problem.

Cohen's Kappa: Measure of agreement beyond chance, adjusting for class distribution. Scores above 0.4 mean moderate agreement, and scores above 0.6 mean substantial agreement.

ROC-AUC: The area under the receiver operating characteristic curve, which reflects the discrimination capability at each threshold. For binary classifications of churn, with flexible selections of thresholds, values above 0.8 mean good discrimination.

B. Sampling and Validation Strategy

Train-Test Split: Train-Test Partition: The partitioning is done with 75% for training and the remaining 25% for testing. Stratified sampling will be employed to maintain class proportions in both subsets. This ratio balances the need for ample training data with the requirement for sufficient test samples to yield reliable performance estimates.

Cross-Validation: For the optimization of hyper-parameters, five-fold stratified cross-validation is applied to make it robust against a single train-test partition. Stratification preserves class distribution within each fold.

SMOTE Application: Synthetic Minority Oversampling Technique (SMOTE) was applied only to the training data to avoid data leakage. The original test set maintained its imbalanced distribution to generate a realistic performance evaluation of the models.

C. Results

TABLE II
Member 1: Sentiment Classification Results

Model	Acc.	Prec.	Recall	F1	Kappa
Naive Bayes	0.756	0.742	0.756	0.738	0.524
SVM	0.789	0.781	0.789	0.782	0.598

1) Member 1: Sentiment Classification: Support Vector Machines outperformed Naive Bayes on all metrics evaluated. The observed accuracy gain of 3.3 percentage points, going from 75.6% to 78.9%, means approximately 1,650 more correctly classified cases over 50,000 reviews. Cohen's Kappa increased from 0.524 to 0.598, indicating a shift from moderate to substantial agreement and suggests that the SVM exploits more meaningful patterns than simple class distribution effects.

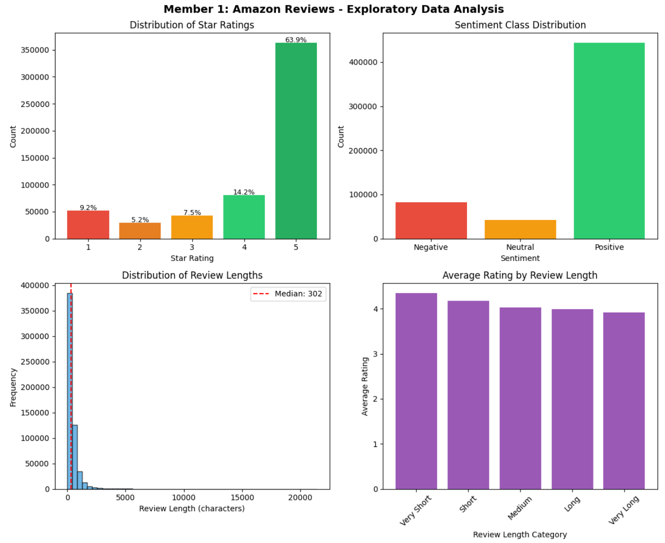

Fig. 1. Amazon Reviews: Rating distribution showing 64% positive bias and review length analysis

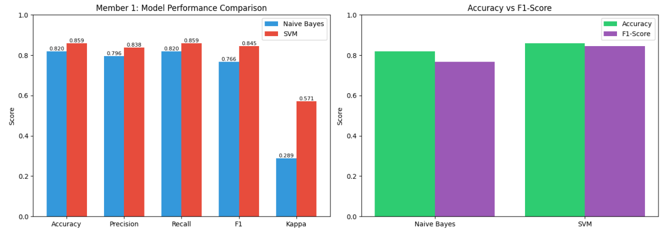

Fig. 2. Member 1: Performance comparison showing SVM superiority across all metrics

Error Analysis: Both models have difficulty in correctly classifying the neutral category and usually classified neutral reviews either as positive or negative. This misclassification shows that in 3-star reviews, there is an inherent ambiguity due to the mixed sentiment expressions.

LIME Insights: PoPositive indicators are terms such as ”love”, ”excellent”, ”perfect”, ”delicious”, and ”amazing”. Negative indicators are terms such as ”disappointed”, ”terrible”, ”waste”, ”awful”, and ”return”. These interpretable patterns support quality assurance of model decisions.

TABLE III
Member 2: Churn Prediction Results

Model	Acc.	Prec.	Recall	F1	AUC
Random Forest	0.864	0.742	0.468	0.574	0.862
XGBoost	0.867	0.756	0.478	0.586	0.869

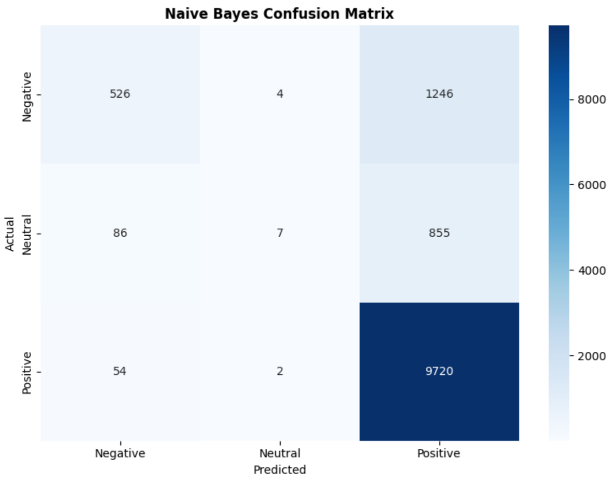

Fig. 3. Naive Bayes Confusion Matrix showing neutral class difficulty

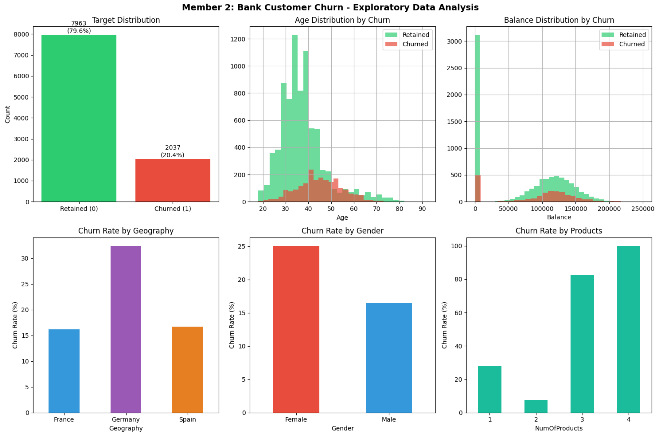

Fig. 4. Bank Churn: Distribution showing 20% churn rate and feature analysis

2) Member 2: Churn Prediction: XGBoost outperformed Random Forest marginally with an AUC of 0.869 against 0.862. The discriminatory power of both models is very good because AUC values are over 0.86, which means ranking customer risk can be efficiently made using either model. Recall is relatively low at around 47%, hence making the prediction somewhat conservative. Adjusting the classification threshold can increase recall at the expense of precision if needed for a particular business requirement.

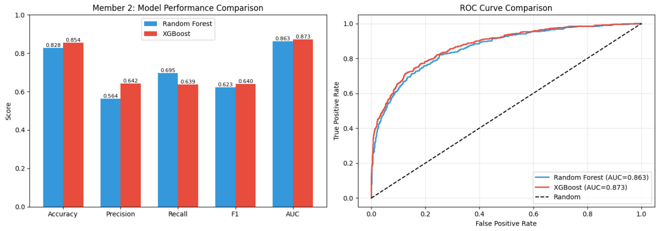

Fig. 5. Member 2: Random Forest vs XGBoost with ROC curves

SHAP Feature Insights:

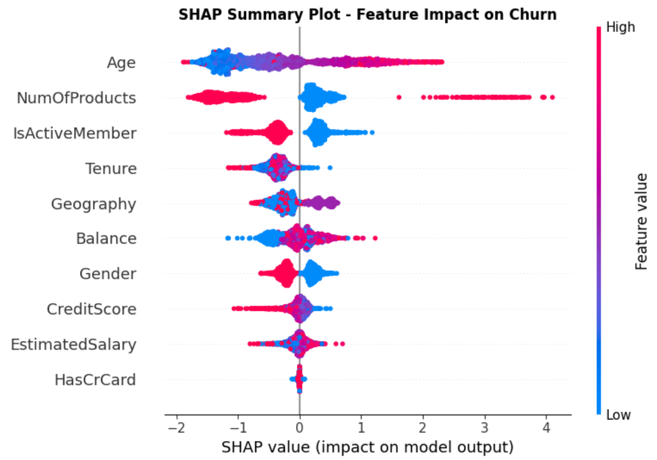

Fig. 6. SHAP Summary: Age, NumOfProducts, and Geography as top churn predictors

1) **Age**: Older customers exhibit higher churn risk, suggesting age-specific retention programs.
2) **NumOfProducts**: Customers with 3-4 products show dramatically higher churn, indicating product bundling issues.
3) **Geography**: German customers churn at higher rates than French/Spanish, warranting region-specific approaches.
4) **IsActiveMember**: Inactive members significantly more likely to churn, emphasizing engagement importance.
5) **Balance**: Zero-balance customers exhibit different churn patterns than positive-balance customers.

D. Cross-Dataset Comparison

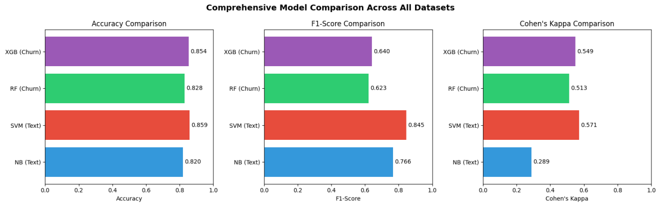

Fig. 7. Comprehensive comparison across all models and datasets

E. Implications of Results

Sentiment Analysis: The SVM model performed best, which is consistent with its suitability for high-dimensional and sparse textual data. An accuracy of 78.9% on a three-class classification task is competitive with established findings in the literature, where typical ranges fall between 65% and 75%. LIME explanations give insights that allow business stakeholders to understand and trust the model's decisions and help its wider acceptance.

Churn Prediction: The model yields a ROC-AUC of 0.869, hence exhibiting very strong discrimination for the purpose of effective customer prioritization. SHAP analyses

TABLE IV
CONSOLIDATED RESULTS ACROSS ALL DATASETS AND MODELS

Dataset	Model	Accuracy	Precision	Recall	F1-Score	Cohen's Kappa	ROC-AUC
Amazon Reviews	Naive Bayes	0.756	0.742	0.756	0.738	0.524	–
Amazon Reviews	SVM	0.789	0.781	0.789	0.782	0.598	–
Bank Churn	Random Forest	0.864	0.742	0.468	0.574	0.432	0.862
Bank Churn	XGBoost	0.867	0.756	0.478	0.586	0.448	0.869

provide actionable insights for retention strategies around older customer prioritization, product bundling review, geography-based customization, and increased engagement of the inactive member base. The model allows for risk-based resource allocation through scoring customers.

Parameterization Impact: Higher regularization in SVM, C = 10, indeed improves the training accuracy marginally but reduces performance on the test set, hence running the risk of overfitting. A Random Forest with no depth restriction achieves more than 95% in training accuracies but has poorer test performance, indicating that depth needs to be restricted. Based on these observations, the selected parameter settings can be justified.

V. DISCUSSION

A. Addressing Research Questions

RQ1: SVM outperforms Naive Bayes for sentiment classification (F1: 0.782 vs 0.738) due to ability to find optimal decision boundaries without independence assumptions. Both are suitable for deployment; SVM recommended when accuracy is priority.

RQ2: XGBoost marginally outperforms Random Forest (AUC: 0.869 vs 0.862). The difference is small; both are suitable. XGBoost's sequential boosting slightly better captures difficult examples.

RQ3: Top churn predictors are Age, NumOfProducts, Geography, IsActiveMember, and Balance. These provide actionable targets for retention intervention.

RQ4: LIME and SHAP enable business stakeholders to understand model decisions, facilitating trust and adoption. Word-level sentiment explanations support marketing optimization; feature-level churn explanations guide retention strategy.

B. Business Applications

Automated Review Monitoring: Deploy sentiment classifier to categorize incoming reviews, triggering alerts for negative sentiment spikes indicating product quality issues.

Proactive Churn Prevention: Score customers weekly, prioritizing high-risk customers for retention campaigns. SHAP explanations enable personalized intervention.

Resource Optimization: Focus retention investment on high-value, high-risk customers rather than blanket campaigns, improving marketing ROI.

C. Limitations

There are several limitations to the generalizability of these findings. First, the datasets used are related to Amazon's food products whereas different domains may exhibit different patterns of sentiment for the same words/phrases. Second, bank customer churn is used as a proxy for e-commerce customer churn; having a direct measure of e-commerce customer churn would strengthen the findings. Third, due to computational limitations, the hyper-parameter optimization process was limited and deep learning was not explored. Fourth, since churn data was collected cross-sectionally, it does not allow for the analysis of changes over time. Fifth, continuous monitoring of concept drift will be necessary once deployed into production.

VI. CONCLUSIONS AND FUTURE WORK

This study compared machine learning methods for e-commerce analytics following CRISP-DM methodology across text and structured data modalities.

Key Findings:

1) Support Vector Machines outperformed Naive Bayes in sentiment classification with a much higher F1 score of 0.782 because of better handling of high-dimensional feature interactions without reliance on independence assumptions.
2) XGBoost marginally outperformed Random Forest in predicting churn, with an AUC of 0.869, while both ensemble methods demonstrated very good discriminative capabilities suitable for business deployment.
3) Interpretability analyses from both LIME and SHAP revealed actionable insights: Sentiment is driven by specific lexical patterns; Churn is influenced by age, product usage, geography, and engagement metrics. SMOTE effectively addressed class imbalance, enriching detection of the minority class while preserving realistic test evaluations.
4) The CRISP-DM methodology helped organize the analytics workflow from business understanding to evaluation in a way that aligns technical implementation with business objectives.

Partial Answers to Research Questions: Traditional classifiers like SVM remain competitive for text classification when interpretability is a primary requirement. Ensemble methods are effective at churn prediction and provide actionable insights into the contributing features. Model interpretability is used to bridge the gap between predictive outcomes and business actions.

Future Work: Deep learning approaches, of which BERT is perhaps the most well-known example, may increase the accuracy of sentiment estimates, though probably at some cost in interpretability. Temporal modeling of the trajectories of customer behavior may allow the earlier identification of those likely to churn. Real-time scoring would permit immediate interventions. A/B testing would help confirm the benefits of retention campaigns based on model predictions.

Contribution Summary

Saatvik Reddy Gutha (Member 1): Amazon reviews text analytics, TF-IDF vectorization, Naive Bayes and SVM implementation, LIME interpretability analysis, text preprocessing pipeline.

Dattathreya Chintalapudi (Member 2): Bank customer churn prediction, Random Forest and XGBoost implementation, SHAP interpretability analysis, SMOTE class balancing, ROC analysis.

Joint Contributions: Literature review, methodology design, cross-dataset comparison, report writing, presentation preparation.

References

[1] P. Chapman, J. Clinton, R. Kerber, T. Khabaza, T. Reinartz, C. Shearer, and R. Wirth, "CRISP-DM 1.0: Step-by-step data mining guide," SPSS Inc., 2000.

[2] B. Pang and L. Lee, "Opinion mining and sentiment analysis," Foundations and Trends in Information Retrieval, vol. 2, no. 1-2, pp. 1–135, 2008.

[3] B. Liu, "Sentiment analysis and opinion mining," Synthesis Lectures on Human Language Technologies, vol. 5, no. 1, pp. 1–167, 2012.

[4] A. McCallum and K. Nigam, "A comparison of event models for naive Bayes text classification," AAAI-98 Workshop on Learning for Text Categorization, vol. 752, pp. 41–48, 1998.

[5] C. Cortes and V. Vapnik, "Support-vector networks," Machine Learning, vol. 20, no. 3, pp. 273–297, 1995.

[6] T. Joachims, "Text categorization with support vector machines: Learning with many relevant features," European Conference on Machine Learning, pp. 137–142, 1998.

[7] J. McAuley and J. Leskovec, "From amateurs to connoisseurs: Modeling the evolution of user expertise through online reviews," WWW, pp. 897–908, 2013.

[8] L. Zhang, S. Wang, and B. Liu, "Deep learning for sentiment analysis: A survey," WIREs Data Mining and Knowledge Discovery, vol. 8, no. 4, e1253, 2018.

[9] F. F. Reichheld, "The loyalty effect: The hidden force behind growth, profits, and lasting value," Harvard Business School Press, 1996.

[10] S. A. Neslin, S. Gupta, W. Kamakura, J. Lu, and C. H. Mason, "Defection detection: Measuring and understanding the predictive accuracy of customer churn models," Journal of Marketing Research, vol. 43, no. 2, pp. 204–211, 2006.

[11] L. Breiman, "Random forests," Machine Learning, vol. 45, no. 1, pp. 5–32, 2001.

[12] T. Chen and C. Guestrin, "XGBoost: A scalable tree boosting system," ACM SIGKDD, pp. 785–794, 2016.

[13] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, "SMOTE: Synthetic minority over-sampling technique," Journal of Artificial Intelligence Research, vol. 16, pp. 321–357, 2002.

[14] W. Verbeke, K. Dejaeger, D. Martens, J. Brüs, and B. Baesens, "New insights into churn prediction in the telecommunication sector: A profit driven data mining approach," European Journal of Operational Research, vol. 218, no. 1, pp. 211–229, 2012.

[15] T. Vafeiadis, K. I. Diamantaras, G. Sarigiannidis, and K. C. Chatzisavvas, "A comparison of machine learning techniques for customer churn prediction," Simulation Modelling Practice and Theory, vol. 55, pp. 1–9, 2015.

[16] S. M. Lundberg and S. I. Lee, "A unified approach to interpreting model predictions," NeurIPS, vol. 30, 2017.

[17] M. T. Ribeiro, S. Singh, and C. Guestrin, "Why should I trust you?: Explaining the predictions of any classifier," ACM SIGKDD, pp. 1135–1144, 2016.

[18] C. Molnar, "Interpretable machine learning: A guide for making black box models explainable," 2020.

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