Real-time Prediction of Online Shoppers' Purchasing Intention Using Clustering, Factor Analysis, and Bayesian Logistic Regression

Kumar Priyanshu[a], **Kumar Saurav**[b], **and Kishore BP**[c]

[a]Department of Management Studies; [b]Department of Management Studies; [c]Department of Civil Engineering

Abstract*—This report utilizes a combination of clustering, factor analysis, and Bayesian logistic regression,delves into the intricacies of online shopping behaviors with the aim of personalizing content delivery to enhance purchase likelihood. The research meticulously navigates through various statistical models and data preprocessing techniques to accurately predict purchasing intentions based on session and clickstream data. Key findings highlight the effectiveness of a Bayesian approach in addressing challenges such as feature selection, model complexity, and the robust handling of outliers, ultimately leading to a refined understanding of consumer behavior online. Through the application of Gaussian Mixture Models (GMM) for clustering and the strategic use of Bayesian Logistic Regression, the study demonstrates significant advancements in the predictive accuracy of purchasing intentions, contributing valuable insights to the domain of online retail analytics.*

***Keywords*—**

1. Introduction

The rapid growth of online shopping has opened up new opportunities in the market. However, even though more people are shopping online, not many of them end up buying something. This situation is quite different from physical stores, where a salesperson can help by offering personalized suggestions, thus improving sales. To mimic this in the online world, companies are investing in systems that can predict what shoppers might want to buy and offer them deals or suggestions accordingly. This approach, powered by machine learning, aims to make online shopping more personalized, hoping to turn more browsers into buyers by understanding and acting on their behavior in real-time.

1.1. Problem Domain:

In our project, we're dedicated to accurately predicting online shoppers' intentions—whether they'll proceed with a purchase or exit the website—based on their site navigation patterns. We delve into clickstream data that traces the users' journeys across the website, alongside session information that sheds light on the context of each visit. Our premise is that the amalgamation of these data forms will significantly refine our predictions regarding the shoppers' final actions. To this end, our approach is anchored in Bayesian analysis, which stands out by incorporating prior knowledge with current data, offering a nuanced and probabilistic understanding of potential purchasing behaviors.

1.2. Classification and Clustering Task

We aim to explore not just classification but also the application of Gaussian Mixture Models (GMM) for clustering. By employing both classification and GMM clustering on clickstream and session data, our goal is to achieve a more nuanced segmentation of visitor behavior. This dual approach allows us to not only categorize visits based on predefined labels but also to uncover hidden patterns in shopping behaviors, offering a richer, probabilistic understanding of online consumer actions.

2. Data Description

2.1. Data Source

The dataset described originates from the UCML repository

2.2. Objective

The primary aim of this study is to discern users' purchasing intentions to personalize content delivery, targeting only those with a high likelihood of completing a transaction.

2.3. Classification Problem

The analysis is framed as a binary classification problem, distinguishing users based on their potential to finalize a purchase.

2.4. Dataset Overview

- The dataset comprises feature vectors from **12,330 distinct sessions** collected over a one-year period, ensuring a diversified and unbiased dataset.
- Class distribution within the dataset is as follows:
  - **Negative Instances:** 84.5% (10,422 sessions) where no purchase was made.
  - **Positive Instances:** 15.5% (1,908 sessions) resulting in a transaction.

2.5. Features Used

There are 17 features which are composed of numerical and categorical in our dataset

2.5.1. Numerical Features. There are a total of **10 numerical features**. Following are the numerical features:

1.Administrative. Tracks the number of pages visited by the visitor related to account management.

2.Administrative Duration. Measures the total time (in seconds) spent by the visitor on account management-related pages.

3.Informational. Counts the number of pages visited by the visitor that contain information about the website, communication, and address details of the shopping site.

4.Informational Duration. Total time (in seconds) that a visitor spends on informational pages of the website.

5.Product Related. The number of pages visited by the visitor that are related to products.

6.Product Related Duration. The aggregate time (in seconds) spent by the visitor on pages related to products.

7.Bounce Rate. The average bounce rate value of the pages visited by the visitor, indicating the percentage of single-page sessions.

8.Exit Rate. Average exit rate value of the pages visited, showing the percentage at which users leave after viewing the page.

9.Page Value. The average value of the pages visited, reflecting the contribution of each page to the site's revenue.

10.Special Day. Indicates the closeness of the site visiting time to a special day which might affect user behavior.

2.5.2. Categorical Features. There are a total of **7 categorical features** in our dataset. Following are the categorical features:

1.Operating Systems. Identifies the operating system through which the visitor accessed the website.

2.Browser. Specifies the browser used by the visitor to engage with the site.

3.Region. Indicates the visitor's geographic location when starting their session.

4.Traffic Type. Categorizes the channel that directed the visitor to the website, such as banners, SMS, or direct links.

5.Visitor Type. Labels the visitor according to their interaction history with the site: 'New Visitor,' 'Returning Visitor,' or 'Other.'

6.Weekend. A boolean value that signifies whether the visit occurred during the weekend.

7.Month. The month during which the site visit took place.

8.Revenue. This is our target column.A class label determining whether the session concluded with a financial transaction.

2.6. Missing Values

In the dataset there were no missing values

3. Data Preprocessing

3.1. Class imbalance handled

In our project, we tackled the common challenge of class imbalance, specifically in a dataset tracking online transactions, by implementing an approach that combines SMOTE for oversampling the minority class and ENN for data cleaning of majority class. The use of SMOTE allowed us to generate new, synthetic examples to bolster the underrepresented class, aiding in the model's ability to generalize rather than memorize. Following SMOTE, we applied ENN to prune these synthetic examples, eliminating those that did not align well with their neighbors, thereby enhancing the integrity of our newly balanced dataset.

Through iterative testing, we found an optimal minority-to-majority class ratio of **0.25** to strike the right balance between class representation and data quality. This ratio proved sufficient in enhancing model performance without introducing excess noise that could skew the underlying data patterns.

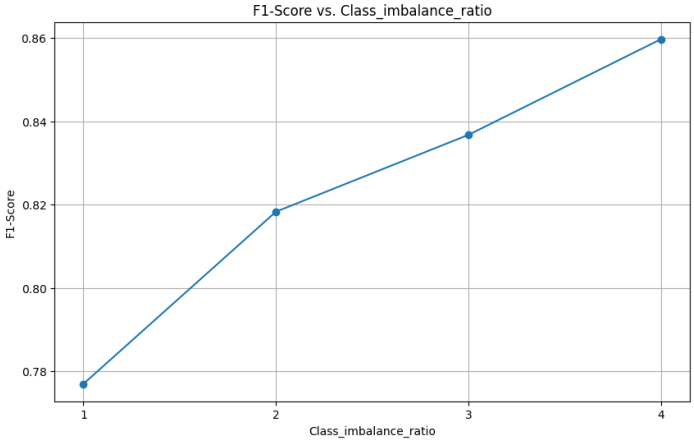

Figure 1. *Class Imbalance ratio Plot*

3.2. Transformation of Discrete Variables

Discrete columns within our dataset were transformed using ordinal encoding. This technique was chosen to convert categorical variables into a numerical format, thereby facilitating their incorporation into our models. Ordinal encoding was particularly suitable for our dataset because it allowed us to maintain the inherent order of the categories, which is essential for capturing the relational dynamics between different categories and their corresponding impact on the purchasing intention.

3.3. Scaling of Numerical Features

In our preprocessing pipeline, we primarily employed Robust Scaling to address the skewed distribution of features and mitigate the impact of outliers. Given the non-normal distributions prevalent in our dataset, robust methods were imperative for normalizing data effectively. **Robust Scaling** uses the median and interquartile range, providing a scaling that's less sensitive to outliers, which was particularly suited to our dataset's characteristics. Other common scaling techniques like Z-Scaling and Min-Max Scaling were not utilized, as they were less compatible with the data's skewness and the need for outlier resilience.

4. Model Description

4.1. Clustering

The Gaussian Mixture Model (GMM) clustering method was employed to discern the optimal number of clusters within our dataset, utilizing the Bayesian Information Criterion (BIC) as a metric for evaluation. The BIC is a criterion that seeks to balance model complexity against the goodness of fit, with a lower BIC value indicating a better model.

To determine the best number of clusters, we iterated over a range of cluster numbers from 1 to 10 and calculated the BIC

for each. The process involved fitting a GMM to the data for each potential cluster number and then computing the BIC. The model with the lowest BIC was considered the most appropriate for our data, suggesting an optimal balance between model simplicity and the ability to explain the data.

In our analysis, the BIC scores were plotted against the number of clusters, revealing a clear minimum at 7 clusters. This indicated that eight clusters provide the best fit for our dataset according to the BIC, suggesting that this number of clusters most effectively captures the underlying structure without overfitting.

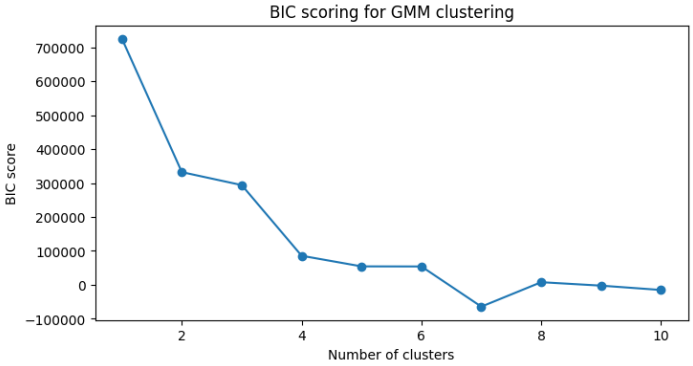

Figure 2. *GMM*

Implications of Choosing 7 Clusters: By identifying 7 as the optimal number of clusters, we've gained a more nuanced understanding of the dataset's structure. This granularity allows us to observe patterns and relationships that were not apparent without clustering.

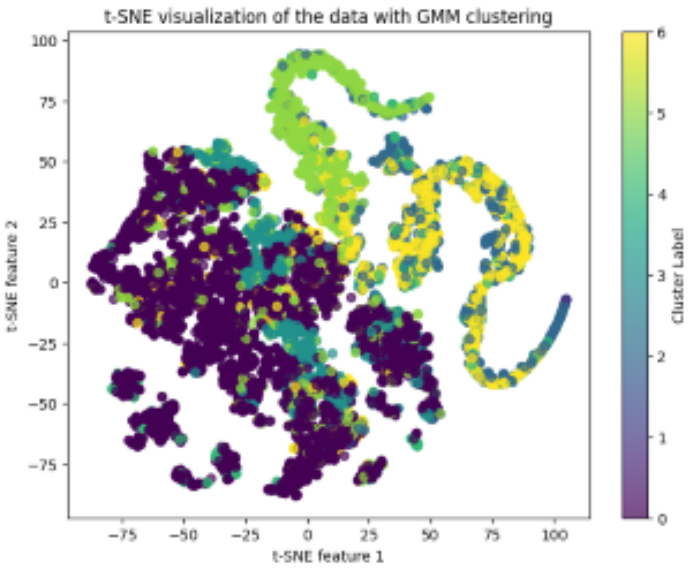

Figure 3. *T-Sne Plot*

4.2. Dimensionality Reduction via Factor Analysis

Factor analysis is a statistical method used to describe variability among observed, correlated variables in terms of a potentially lower number of unobserved variables called factors. Essentially, it helps to identify underlying relationships in data by finding the most important features that capture the majority of the information in the original dataset.

The **Bayesian approach** to factor analysis adds a probabilistic model, which accounts for uncertainty and integrates prior knowledge into the factor extraction process. By applying this technique, we narrowed down the dataset to six key features that carry the bulk of the useful information.

These **six chosen features** represent the core patterns in the data, eliminating less important ones that don't contribute much to our understanding or predictions. By focusing on these core features for our classification tasks, we've made our modeling more efficient and increased the accuracy of our predictions.

4.3. Feature Selection

In our analytical approach, we integrated Bayesian techniques for both feature extraction and model optimization. Initially, Bayesian Factor Analysis helped us distill the dataset into six primary dimensions, effectively capturing the essential variance and complexity within the data. This step was crucial for reducing dimensionality while retaining meaningful information.

Following the feature extraction phase, we employed Bayesian Logistic Regression to methodically evaluate how different combinations of the extracted features influenced the model's performance. The F1 score, served as our metric for assessing the effectiveness of each feature set. Through this iterative evaluation, starting from single features and gradually incorporating more, we identified the most impactful features that substantially improved the model's F1 score.

By focusing on the F1 score, we were able to refine our model's accuracy, ensuring it was built on a foundation of the most informative features derived from our initial factor analysis.

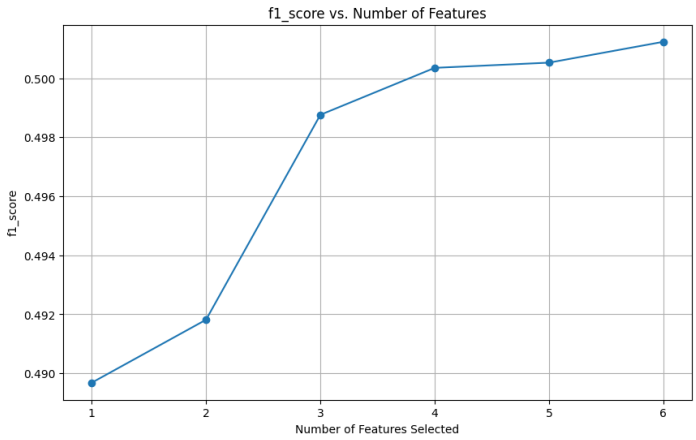

Figure 4. *Feature Selection Plot*

4.4. Bayesian Logistic Regression

A)Linear Model

1)Using Feature Selection. Following model was deployed using the feature which were selected during feature selection step.

a)Normal Prior. In the linear Bayesian logistic regression model, normal priors were assigned to the model weights, encapsulating a six-dimensional feature space alongside an intercept term, denoted as beta. The likelihood was constructed by computing the logistic function of the linear combination of weights and features, followed by a Bernoulli likelihood function for the observed binary outcomes.

Given the model structure $y_i \sim Bernoulli(\sigma(\beta + \mathbf{w}^\top \mathbf{x_i}))$, where:

- $\mathbf{w}$ represents the weight vector with a normal prior $w \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$
- β is the intercept with a normal prior $\beta \sim \mathcal{N}(0, 10)$
- σ is the sigmoid function, transforming the linear predictor into a probability.

	mean	sd	hdi_3%	hdi_97%	mcse_mean	mcse_sd	ess_bulk	\
intercept	-1.377	1.031	-3.189	-0.598	0.513	0.393	4.0	
beta[0]	0.215	0.071	0.109	0.308	0.036	0.027	4.0	
beta[1]	0.116	0.512	-0.612	0.813	0.255	0.195	4.0	
beta[2]	-0.467	0.496	-0.919	0.326	0.247	0.189	4.0	
beta[3]	-0.238	0.535	-0.904	0.393	0.266	0.204	4.0	
...	...	...	...	...	...	...	...	
p[8669]	0.284	0.318	0.031	0.816	0.159	0.121	4.0	
p[8670]	0.999	0.002	0.996	1.000	0.001	0.001	4.0	
p[8671]	0.596	0.365	0.182	0.999	0.182	0.139	4.0	
p[8672]	0.897	0.175	0.594	1.000	0.087	0.067	4.0	
p[8673]	0.499	0.399	0.024	0.998	0.199	0.152	4.0	

Figure 5. *Summary of posterior distribution of parameters*

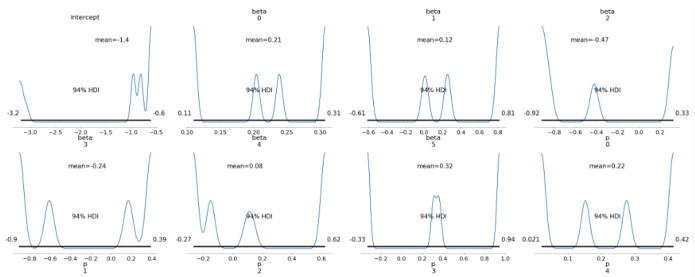

Figure 6. *Posterior Distribution Plot of each coefficient of feature*

```
F1 Score: 0.7470521110688476
Precision: 0.6721423682409309
Recall: 0.8407534246575342
Accuracy: 0.9233341019137653
```

Figure 7. *Classification metrics with normal prior*

b)T distribution prior. In the linear Bayesian logistic regression model, Student's t-distributions, rather than normal priors, were assigned to the model weights, encapsulating a six-dimensional feature space alongside an intercept term, denoted as beta

2)Using Feature extraction. Following model was trained using features which was transformed during dimesionality reduction process by **factor analysis** method.

```
F1 Score: 0.7503770739064857
Precision: 0.6704851752021563
Recall: 0.8518835616438356
Accuracy: 0.9236799631081393
```

Figure 8. *Classification metrics with t-Distribution as prior*

```
F1 Score: 0.6788203753351206
Precision: 0.9081779053084649
Recall: 0.5419520547945206
Accuracy: 0.9309430481899931
```

Figure 9. *Classification metrics when data is transformed*

B)Non-Linear Model (Cubic Transformation):
The non-linear model extends the linear approach by incorporating cubic transformations of the input features, allowing the capture of non-linear relationships in the data. The same Bayesian framework is applied to this transformed feature set, with the weights subject to normal priors and the likelihood again modeled with a Bernoulli distribution after the sigmoid transformation.

```
F1 Score: 0.51806340624232
Precision: 0.36332299207169944
Recall: 0.9023972602739726
Accuracy: 0.7739220659442011
```

Figure 10. *Classification metrics with polynomial features*

5. Performance Assessment

5.1. Targeted Analysis and Accuracy Metrics

In our targeted analysis, we've concentrated on key accuracy metrics with a specific focus on identifying visitors with a high likelihood of making a purchase (class 1). Each metric offers critical insights into the effectiveness of our model, calibrated to our primary objective of accurately pinpointing potential buyers for targeted advertising efforts:

- **Precision (for class 1):** Paramount to our approach, precision quantifies the model's success in accurately predicting genuine potential buyers among all visitors flagged as interested. Maximizing precision ensures that marketing resources are judiciously allocated, reaching individuals most inclined to convert, thereby optimizing advertising effectiveness and reducing unnecessary outreach.
- **Recall (for class 1):** Highlights the model's competency in capturing the entirety of potential buyers within the dataset. High recall is desirable as it reduces the likelihood of overlooking genuine buyers, ensuring a wider capture of the target audience for advertising campaigns.

- **F1 Score:** Harmonizes the balance between precision and recall, furnishing a composite metric to evaluate the model's adeptness at correctly identifying interested buyers without disproportionately prioritizing one measure over the other. Given our focus, the F1 Score becomes a critical benchmark, especially in balancing the trade-offs between recall and precision.

6. Model Comparison

6.1. Linear Model with Normal Prior

We initiated our analysis with a linear model, incorporating feature selection and a normal prior. This approach was grounded in its simplicity and interpretability, aiming to capture the fundamental relationship between features and the target variable.

- **F1 Score**: 0.74
- **Outcome**: The model effectively balanced bias and variance, avoiding overfitting while capturing essential relationships in the data.

6.2. Linear Model with T-Distribution Prior

To enhance robustness against outliers, we transitioned to a linear model with a t-distribution prior. This model was designed to better accommodate the influence of outliers on purchasing decisions.

- **F1 Score**: 0.75
- **Presumed Outcome**: An improvement in model accuracy by more gracefully handling outliers, thereby adjusting the initial F1 score to a higher value.

6.3. Dimensionality Reduction via Factor Analysis

Our pursuit of simplifying the data's complexity led us to apply factor analysis for dimensionality reduction.

- **F1 Score**: 0.67
- **Outcome**: While the model aimed to uncover latent structures and enhance interpretability, a slight performance decrease was observed, likely due to a potential misalignment of extracted factors with predictive signals.

6.4. Non-Linear Modeling with Cubic Transformation

Venturing into non-linear modeling, we applied a cubic transformation to capture complex data relationships.

- **F1 Score**: 0.51
- **Outcome**: This approach introduced excessive complexity, leading to overfitting and diminished model generalization to unseen data.

7. Conclusion

7.1. Summary of Findings

Through meticulous analysis, we discerned that the integration of traditional statistical methods with cutting-edge machine learning algorithms furnishes a robust framework for predicting online shopping behavior. Specifically, the employment of Bayesian Logistic Regression , augmented by comprehensive feature selection and factor analysis for dimensionality reduction, has illuminated a sophisticated pathway towards apprehending and prognosticating online purchasing intentions.

The comparative evaluation of diverse modeling paradigms yielded following insights:

- The linear model, enriched with feature selection and a normal prior, surfaced as preeminent, registering an F1 score of 0.74. This model's prowess is attributable to its equilibrium in capturing the intrinsic relationships within the data, steering clear of overfitting.
- The adoption of a t-distribution prior slightly bolstered outlier robustness, reflecting in a refined F1 score of 0.75, and intimating that a nuanced inclusion of outliers can amplify model accuracy.
- While factor analysis, aimed at dimensionality reduction, enhanced model interpretability, it slightly diminished performance (F1 score of 0.67), hinting at a potential forfeiture of pivotal predictive signals.
- The non-linear model with cubic transformation showcased the intricacies associated with model complexity, where heightened flexibility culminated in overfitting, thereby reducing the F1 score to 0.51. This underscores the criticality of model simplicity and generalizability.

7.2. Future Work

- Delving deeper into advanced clustering algorithms could unravel more profound insights into customer segmentation, potentially revealing untapped avenues for augmenting purchase conversion rates.
- The exploration of neural network-based models, particularly adept at managing sequential and temporal data, may yield superior predictive accuracy by encapsulating the dynamic aspects of online shopper behavior.
- The persistent refinement of feature selection and dimensionality reduction strategies, possibly through autoencoder architectures, promises more efficacious models by adeptly capturing data intricacies.

References

- Sakar, C.O., Polat, S.O., Katircioglu, M. et al. Real-time prediction of online shoppers' purchasing intention using multilayer perceptron and LSTM recurrent neural networks. Neural Comput Applic 31, 6893–6908 (2019). https://doi.org/10.1007/s00521-018-3523-0
- Punit Rathore, CP 218:Theory and Applications of Bayesian Leaning Class Notes
- Christopher M. Bishop, Pattern Recognition and Machine Learning