Customer Purchase Prediction Using Machine Learning

Trevor McGirr
Department of Engineering
University of San Diego
AAI-550-01-FA22
Dr. Ying Lin

Abstract

Customer purchase predictions is valuable for a marketing team to drive sales for a business. In this study, prediction models of customer purchasing behavior and demographics were analyzed to determine the most accurate model. The Extra Tree Regressor model was found to have the highest Coefficient of determination (R^2) of 0.83, indicating its ability to predict the amount of wine a customer would purchase. The most important features in this model were "Number of Store Purchases", "Income", "Number of Catalog Purchases", "Accepting a Coupon by the 5th Period", and "Amount of Meat Purchased". These features can be used by the marketing team to target customers with similar characteristics in order to drive wine sales. Further research will involve using the model to predict wine purchasing behavior and targeting customers accordingly.

1. Introduction

Predicting customer purchasing behavior can provide a business with valuable insights for driving sales and increasing profits. Knowing the factors that influence a customer's decision to buy a certain product allows a business to tailor their marketing strategy to effectively target specific customer demographics. Machine learning algorithms can be used to analyze existing customer data including demographics and purchasing behavior to develop accurate prediction models. In this research article, we will evaluate different regression prediction models and identify the most effective model for predicting customer purchasing behavior.

2. Literature Review

This section will provide a review of the previous research on customer purchase predictions and business implications. This section will also introduce and provide context for the commonly used methods for customer purchase prediction, including various machine learning models that have been used and their efficacy. In addition, the challenges and limitations of model selection and evaluation metrics used to measure performance will be discussed. The literature review section will provide a foundation for the context of the study design and the purpose of finding the optimal prediction algorithm.

2.1 Customer Purchase Prediction

Customer purchase prediction provides a company with the ability to conserve marketing resources and drive sales and profits by targeting specific demographics and advertising specific products. The customer acquisition cost

far outweighs the cost to retain an existing customer. A cost-effective approach to driving sales would include increasing marketing to existing customers for products that align with their purchasing profile. Martinez et al (2020) describe the benefit of purchasing predictions as one of the key drivers to allocate sales and marketing department resources efficiently for both sales forecasting and inventory management.

The application of machine learning has been applied to a variety of tasks related to customer price prediction, including regression, classification, and clustering. The application of regression algorithms provides the prediction of consumers' purchase intentions (Chen et al 2021). Another approach to customer price prediction is the use of classification algorithms, such as logistic regression, naive Bayes, and k-nearest-neighbors. The classification algorithms help predict the likelihood of a customer belonging to a particular category based on their characteristics and behavior. The clustering algorithms, such as k-means, group customers into distinct clusters based on their characteristics and purchase behavior. These machine learning algorithms are trained on previous customer data, including customer demographics, purchase history, and product categories, in order to define the relationship between the variables and customer purchase tendencies.

2.2 Machine Learning Models

Regression models are primarily used for the prediction of a continuous/real outcome, to quantify the relationship between the input variables (Maulud 2020). The linear regression model assumes a linear relationship between the predicted input variables and the predicted outcome. The linear regression model can include a single independent variable, Simple Linear Regression, or multiple inputs, Multivariate Linear Regression (Maulud 2020). Due to the assumption of a linear relationship between independent and dependent variables, the accuracy of a linear regression model may decrease for complex non-linear relationships.

Support Vector Machine models will separate two classes by fitting a hyperplane between them that maximally separates the data points belonging to different classes (Seippel 2018). The SVM model can be effective when the data is not linearly separable by using a kernel trick to project the data into a higher-dimensional space where it can be linearly separated. Due to the sensitivity to kernel choice and other hyperparameters, there must be a careful tuning and selection process for accurate results.

Decision Trees Models recursively partition data into smaller subsets based on a certain variable to create the most homogeneous subgroups (Seippel 2018). Through each separation, the model selects the features and threshold values that will

produce the purest subset, which is determined by the Gini impurity or entropy. The result of the decision tree will be a tree-like structure of decision rules that can be used to make new predictions. However, due to the sensitivity of the model and the impact of the features selected, the model is prone to overfitting and may produce inaccurate results.

2.3 Research Purpose

The application of an appropriate machine learning model is paramount to producing accurate and relative results. With each model having its own strengths and limitations, the appropriate model for a given situation may vary depending on the specific data and goals of the business. When attempting to predict a customer's purchasing behavior, there are several considerations when selecting the desired outcome and what data points are available to train the model. Some factors that may be important for predicting consumer behavior include demographics, income, and purchasing history.

Previous research has also explored the use of different evaluation metrics, such as the coefficient of determination (R^2) and mean squared error, to measure the performance of prediction models. This study will aim to find the most accurate model for predicting customer purchasing behavior and provide an evaluation and application of the provided model.

3. Methodology

3.1 Data Preparation

The dataset used for this study was provided by Kaggle "https://www.kaggle.com/datasets/whenamancodes/customer-personality-analysis". The initial data exploration was conducted using the sklearn library to understand the columns, data points, and data types within the dataset. The dataset included 29 columns, 2240 entries, and datatypes consisting of "float64", "int64", and "Object" types. There were 24 missing values in the "income" column, which were replaced with the mean of the column. There was one outlier displayed in the "income" column as well, which was excluded for training the model training data frame. After the missing data was processed, the categorical data "Education" and "Marital Status" was transformed using the Label Encoder from sklearn, and the unnecessary identification columns were dropped.

With the processed data points, the data frame was then split into "X" and "y" test and train sets, consisting of a data frame of features without the Quantity of Wine Amount ("MntWines") for the independent "X" and the data frame consisting of only the Quantity of Wine Amount ("MntWines") for dependent "y". The training and test sets were then split using the sklearn "train_test_split" function from sklearn with an 80% (Train) and 20% (Test) split.

The split training set “X_train” and “y_train” were then scaled using the “Standard Scaler” method from sklearn to transform the data into a standardized value, creating a mean of 0 and a standard deviation of 1 for each feature.

3.2 Model Training

With the training data and features properly chosen, encoded, and scaled, a variety of regression models were then trained, evaluated, and compared for future hypertuning. Each model was fit with the provided testing dataset and evaluated using a “Mean Squared Error” and “Coefficient of Determination”, or R-Squared score.

The Linear regression model assumes a linear relationship between the input features. The Random Forest model is an ensemble model that trains multiple decision trees on random subsets and averages their predictions. The Support Vector model uses support vector machines to find the optimal hyperplane for making predictions. The Decision Tree regression model makes predictions by splitting the data into increasingly pure subsets based on input features. The K-Nearest Neighbors regression model makes predictions based on the average of the target values of the nearest selected neighbors, which was “5” for this model. The XGBoost regression model is an ensemble model that trains multiple decision trees using the gradient boosting algorithm. The Gradient Boosting regression model is an ensemble model that trains multiple decision trees, where each decision tree is trained to correct the mistakes of the previous tree. The AdaBoost regression model is an ensemble model that trains multiple decision trees, where each tree is given a weight based on its performance. The Bagging regression model is an ensemble model that companies the same base model on different subsets of the data and averages their predictions. The Extra Trees regression is an ensemble model that trains multiple decision trees on random subsets of the data and averages their predictions. The Stochastic Gradient Boosting model is an ensemble model that trains multiple decision trees using the stochastic gradient boosting algorithm. The Voting regression model is an ensemble model that trains multiple models and makes predictions based on the majority vote of the individual model predictions.

4. Results

4.1 Model Evaluation

With each of the models being fit to the same training data and being evaluated by their R-Squared score, the highest performing model was selected for hypertuning and refitting. The R-Squared score of each model is provided in Figure 1. Below.

The Extra Trees regression model performed the best of all of the models and the GridSearchCV method from sklearn was used to find the

optimal parameters to be used to fit the model. The Extra Trees Regression model was then re-fit using the optimal parameters attained from the GridSearchCV. The score of the re-fitted Extra Decision Trees model remains at “0.83” for the R-Squared score and proved to be the best performing model for the Customer Purchase Prediction dataset for the “Quantity of Wines” prediction.

4.2 Feature Importance

Using the Extra Trees regression model and optimal parameters provided by the GridSearchCV, the features were ranked by importance in their contribution to the model and provided in Figure 2.

5. Conclusion

The top model based on the highest R-Square score was the Extra Decision Trees with a score of “0.83”, which indicates a relatively low error and a good ability to predict the target variable of “Quantity of Wine” a customer may purchase. The top 5 features in terms of importance are NumStorePurchases, Income, NumCatalogPurchases, AcceptedCmp5, and MntMeatProducts. These features have importance values of 0.18, 0.15, 0.15, 0.11, and 0.08 respectively, indicating that they have a significant impact on the performance of the model.

The NumStorePurchases refers to the number of store purchases made by a customer, and its high importance suggests that this feature is strongly related to the target variable of “Quantity of Wine” purchased. Income, which is the customer's income level, is also an important predictor, likely because higher income is associated with certain behaviors or characteristics that are relevant to the target variable. NumCatalogPurchases, which is the number of catalog purchases made by an individual, is also an important feature, indicating that this behavior is related to the target variable as well. AcceptedCmp5, which is the individual's participation in accepting a coupon by the fifth round of promotions indicates the individual's use of coupons. The NumMeatProducts, which is the number of meat products purchased by the individual also displays a relation to the purchasing amount of wine.

The next steps for further research could include conducting additional analysis to validate the results and identify potential limitations of the model or using Neural Networks. In addition, it would be beneficial to explore other potential features and variables that may have an impact on the target variable of "Quantity of Wine" purchased. Overall, further research could help to improve the accuracy and predictive ability of the model and provide additional insights into the factors that influence the purchasing behavior of customers.

Figure 1: Model R-Squared Score

Model	R-Squared Score
Linear Regression	0.71
Random Forest Regression	0.82
Support Vector Regression	0.18
Decision Tree Regression	0.60
K-Nearest Neighbors	0.69
XGBoost Regression	0.82
Gradient Boosting Regression	0.81
AdaBoost Regression	0.62
Bagging Regression	0.80
Extra Tree Regression	0.83
Stochastic Gradient Boosting Regression	0.81
Voting Regression	0.81

Figure 2: Feature Importance

Variable	Importance
NumStorePurchases	0.18
Income	0.15
NumCatalogPurchases	0.15
AcceptedCmp5	0.11
MntMeatProducts	0.08
Kidhome	0.06
NumWebPurchases	0.05
NumWebVisitsMonth	0.05
Education	0.03
MntFishProducts	0.02
AcceptedCmp4	0.02
Year_Birth	0.01
Marital_Status	0.01
Teenhome	0.01
Recency	0.01
MntFruits	0.01
MntSweetProducts	0.01
MntGoldProds	0.01
NumDealsPurchases	0.01
AcceptedCmp3	0.01
AcceptedCmp1	0.01
Response	0.01
AcceptedCmp2	0.0
Complain	0.0
Z_CostContact	0.0
Z_Revenue	0.0

References

Chen, S., Wang, X., Zhang, H., & Wang, J. (2021). Customer purchase prediction from the perspective of imbalanced data: A machine learning framework based on factorization machine. *Expert Systems with Applications,* 173, 114756. https://doi.org/10.1016/j.eswa.2021.114756

Maulud, D., Abdulazeez, A. (2020). “A review on linear regression comprehensive in machine learning. *Journal of Applied Science and Technology Trends*, vol. 1, no. 4, pp. 140–147

Martínez, A., Schmuck, C., Pereverzyev, S., Pirker, C., & Haltmeier, M. (2020). A machine learning framework for customer purchase prediction in the non-contractual setting. *European Journal of Operational Research,* 281(3), 588-596. https://doi.org/10.1016/j.ejor.2018.04.034

Seippel, H. (2018). Customer purchase prediction through machine learning. MS thesis. The University of Twente. https://essay.utwente.nl/74808/1/seippel_MA_eemcs.pdf