Using Data Mining to predict Customer′s Net Profit

Prof:Seif Al-Din FattouhOsman

Mohammed Mahjoub Ibrahim Mohamed

NeelainUniversity

Abstract:

The importance of decision-making in business depends entirely on the extent of definite knowledge of achieving profits while reducing losses as much as possible based on potential expectations and based on real data, and accordingly this study aims to achieve the main objective of the message is to predict the net profit for the customer, the research gained its importance To provide highly accurate and reliable analysis tools to obtain the best results and expectations when studying entering new markets or business operations, a mini-model was developed in Python language based on the multiple linear regression equation to predict the expected net profit from the customer at each visit to him, the results that reach Her research is to predict the net profit of the customer based on some actively neglected traits.

Keywords:multilinear- regression, customer net profit, prediction.

إستخدام تنقيب البيانات للتنبؤ بصافي الربح للزبون

أ.د: سيف الدين فتوح عثمان - قسم نظم المعلومات – الكلية الاماراتية – السودان

أ. محمد محجوب إبراهيم محمد – طالب دكتوراه-جامعة النيلين

المستخلص:

إن أهمية اتخاذ القرار في الاعمال التجارية يعتمد كليا على مدى المعرفة الاكيدة بتحقيق الارباح مع تقليل الخسائر بقدرالامكان بناء على التوقعات المحتملة والمبنية على بيانات حقيقية, وبناء على ذلك تهدف هذه الدراسة الى تحقيق الهدف الرئيس هو التنبؤ بصافي الربح للزبون، اكتسب البحث اهميتة لتقديمة ادوات تحليل ذات دقة عالية ويعتمد عليها في الحصول على افضل النتائج والتوقعات عند دراسة الدخول في اسواق اوعمليات تجارية جديدة، تم تطوير نموذج مصغر بلغة بايثون مبني على معادلة الانحدار الخطي المتعدد للتنبؤ بصافي الربح المتوقع من الزبون عند كل زيارة له،النتائج التي توصل لها البحث هي اللتنبؤ بصافي ربح الزبون بناءعلى بعض الصفات المهملة بشكل فاعل.

الكلمات المفتاحية: الانحدارمتعدد الخطوط ،صافي ربح العميل،التنبؤ

1. Introduction

Predicting future customer behavior provides key information for efficiently directing resources at the sales and marketing departments(1). (١)Customer satisfaction has become one of the most important factors for any institution. With the spread of electronic services customer awareness has increased, retaining a customer is very important in this competitive market. Firms invest heavily in customer relationship management and sales automation tools to improve sales force(2).

In recent years there are much data as well as many fields of applications where data is being recorded(3). 0.64 of sales managers rated that designing a sale model as one of the top three challenges(4). The dataset was collected fromdata.worldonline repository, the dataset of retail sales transactionscontaining 12 attributes and 34,867 records as follows: visit_date, customer_age,customer_gender, country, state, production_category, sub_category, quantity, unit_cost, unit_price, total_cost, and revenue. Between the years 2015 and 2016.multilinear regression was used as a method to predict customer net profit for each visit, the model would enable sales management to action for the customer in a short period of time such as providing a discount for the customer or any future offers.

The rest of this paper is organized from sections 2 to 12 as follows: related works, customer experience, data collection, data cleaning and transformation, methodology, exploratory data analysis, model development, results, discussion, and future work respectively.

2. Related Studies

In paper(5), the authors built a model that could predict a customer's likelihood of subscribing to products, offers, and other packages for the bank. They applied four classification algorithms: multilayer perception neural network, decision tree (C4.5), logistic regression, and random forest, they could determine key features and they found that customers with more call duration were more likely to subscribe to the tern deposit. The authors in paper(1), developed an advanced analytics model that could predict future customer behavior, to determine whether a customer is going to make a purchase within a certain time frame in the near future. They used the following algorithms: logistic regression, extreme learning machine, gradient tree boosting. In paper (6), a new manufacturing resource planning model was developed to predict the manufacturing resource requirements from customer demands by reviewing the relations between customer demands and manufacturing resource requirements, the sales data was collected from all levels of the company departments, branches, markets, and requirements of different production lines were studied. The following algorithms were used clustering and linear regression. In research(7), the authors designed a multichannel customer segmentation model to predict segment membership, they took into account channels used for information search and product purchase. Furthermore, they included after-sales services and utilized self-report behavior. The paper(8), introduced an approach that utilized heterogeneous social networks to improve the effectiveness of offline sales and identify potential enterprise customersfor offline sales. In paper(9), the authors proved that customer orientation did have a significant direct effect on longitudinal sales performance, while there was no direct effect on cross-sectional sales. In paper(10), they proposed a framework to predict the revisit intention of first-time customers, they used Wi-Fi signals captured by in-store sensors. They proved that mobility features were important even with a few numbers of records. In paper(11), the authors investigated the probability of increasing sales based on the similarity between employees and customers. In the study(12), the authors developed three approaches for location and next location prediction based on mobile location data to support delivery planning, in addition, this will decrease last-mile delivery costs and boost customer satisfaction.

3. Customer Experience

The author(13), addressed the identification of the most profitable customers' groups using cluster analysis. Building a long-term loyal customer is a key competitive factor. Salespersons play a significant role in creating customer value(14). In paper(15), the authors used satisfaction to predict company growth rates. Building a prediction machine learning model of remanufactured products using online market factors(16).customer attitudes and customer behavior has been gaining significant attention with the growing recognition that customers are market-based assets(17).

4. Multiple linear regression analysis

Regression analysis is a statistical technique for estimating the relationship between two or more variables. Regression answers questions like are there any relationship among the dependent and independent variables? If there is the power of the relation? Is it possible to make a future prediction based on the dependent variable? If certain conditionswere controlled, what influences does a special variable or a group of variables have over another variable or variables(18)? Regression analysis is often used in different fields and applications(19).

Linear regression assumes a linear relationship between input variables and the output variable. When more than one variable is used it is called

multiple linear regression (20).Linear models are simple but usually provide a clear description of how input variables contribute to the output.

Y= $\beta\theta + \beta 1X1 + \beta 2X2 + \cdots + \beta pXp + \epsilon$ $\beta\theta + \beta 1X1 + \beta 2X2 + \cdots + \beta pXp + \epsilon$

Where Y is the output variable, $\beta i\beta i$ regression coefficient X j are the variables, $\epsilon\epsilon$ is the intercept. The parameters are computed and estimated by using statistical software. It was used to predict the net profit which the expat is expected to stay abroad. (20). Mentioned that the accuracy of the statistical prediction is related to the amount of data, the more data is available it would be useful for prediction. (21).multiple linear regression models describe how a single response variable depends linearly on a number of predictor variables.

4.1. Estimates of the Model Parameters

The estimates of the β coefficients are the values that minimize the sum of squared errors for the sample. The exact formula for this is given in the next section on matrix notation.

The letter b is used to represent a sample estimate of a β coefficient. Thus $b0$ is the sample estimate of $\beta0$,$b1$ is the sample estimate of $\beta1$, and so on.

In the case of two predictors, the estimated regression equation yields a plane (as opposed to a line in the simple linear regression setting). For more than two predictors, the estimated regression equation yields a hyper plane.

4.2. Interpretation of the Model Parameters

- Each β coefficient represents the change in the mean response, E(y), per unit increase in the associated predictor variable when all the other predictors are held constant.

- For example, $\beta 1$ represents the change in the mean response, E(y), per unit increase in $x1$ when $x2$, $x3$, ...,$xp-1$ are held constant.
- The intercept term, $\beta 0$, represents the mean response, E(y), when all the predictors $x1$, $x2$, ...,$xp-1$, are all zero (which may or may not have any practical meaning).

5. Data Collection

The dataset was collected from data.world online repository, the dataset of international retail sales transactions containing 12 attributes and 34,867 records as follows: visit_date, customer_age, customer_gender, country, state, production_category, sub_category, quantity, unit_cost, unit_price, total_cost, revenue. It was inthe years 2015 and 2016.

6. Data cleaning and Transformation

As linear models work with numerical variables transformation process was needed. Attributes like customer_gender, production_category, sub_category, and country.They were converted into numerical values, before the analysis process. New attributes were generated from the dataset, they were: visit_month, cost_percent, profit_percent, and net_profit.

Table1.Showed the transformationsofattributes.

Attribute	Value
Product_category	1=Accessories 2=Clothing 3=Bikes
country	1=United States 2=France 3=United Kingdom 4=Germany
Age	Continuous

Attribute	Value
Sub_category	1=Tires and Tubes 2=Gloves 3=Helmets 4=Bike Stands 5=Mountain Bikes 6=Hydration packs 7=Jerseys 8= Fenders 9=Cleaners 10=Socks 11=Caps 12=Touring Bikes 13=Bottles and Cages 14=Vests 15=Read Bikes 16=Bikes Roads 17=Shorts
Customer_gender	1=male 2=female

7. Methodology

Multiple linear regression was used as a method for predicting the net profit for the customer in each visit. Input attributes were as follows: visit_month, customer_age, customer_gender, and country. The response variable was net_profitAlinear model was formed using the following formula,

$$Y= \beta\theta + \beta 1X1 + \beta 2X2 + \cdots + \beta pXp + \epsilon$$

$$\beta\theta + \beta 1X1 + \beta 2X2 + \cdots + \beta pXp + \epsilon$$

Where:

Y = net_profit

X1 = visit_month
X2 = customer_age
customer_gender=X3
Bi = Coefficient
E = Intercept

8. Exploratory Data Analysis

Exploratory Data Analysis (EDA) refers to the critical process of performing initial investigations on data so as to discover patterns, spot anomalies, test hypotheses and to check assumptions with the help of summary statistics and graphical representations.

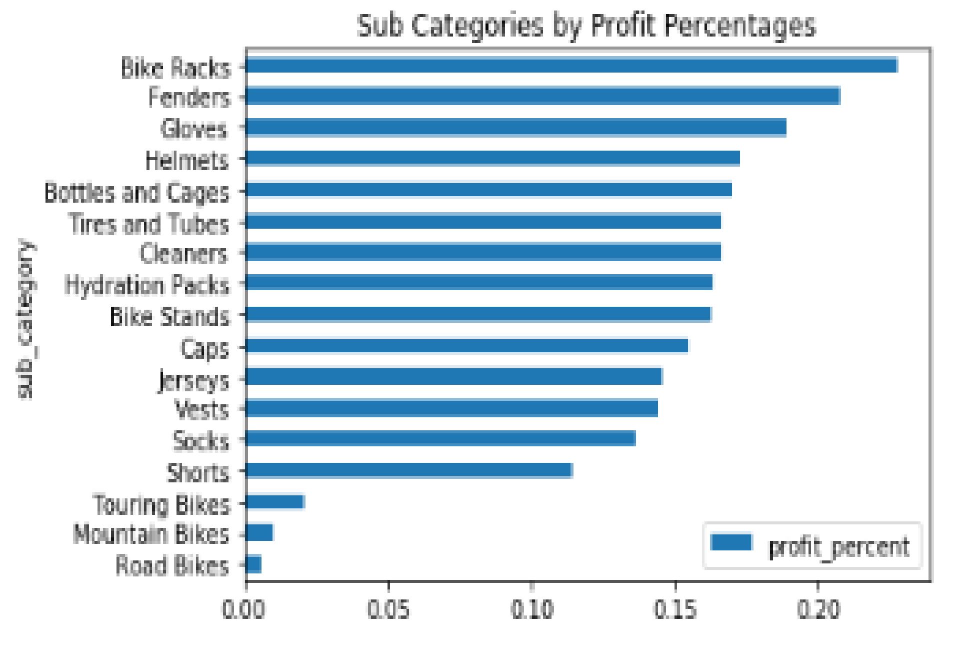

Fig1.Showed subcategories and profit percentages.

The figure above presents the subcategories and their percentages, it is clear that Bike Racks, Fenders, and Gloves were the top three profitable subcategories.

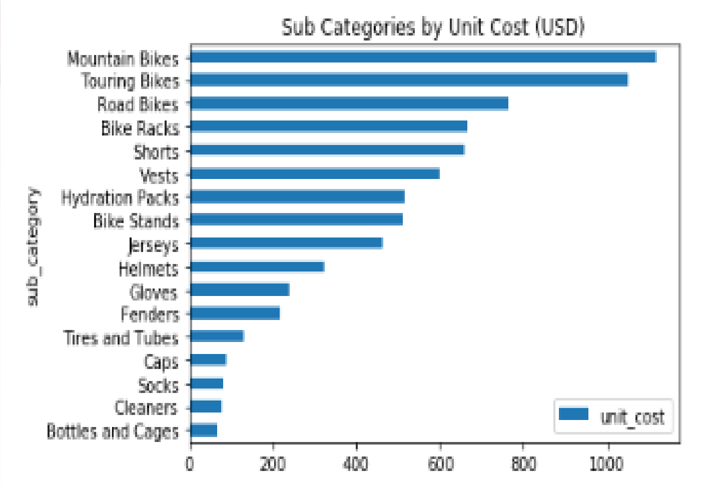

Fig2.Presented subcategory and unit cost.

From the fig above it was found that Mountain Bikes, Touring Bikes, and Roads Bikes were the most three subcategories in terms of cost.Whereas, bottles, cages, cleaners, and socks were the least.

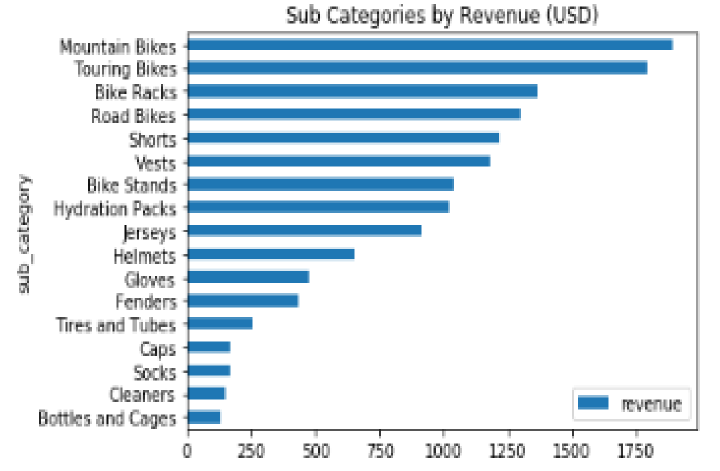

Fig3.Illustrated subcategories and revenue.

According to the fig above, the subcategories mountain bikes, touring bikes were the top two categories that bring revenue. However, socks, cleaners, and bottles and capes were the least.

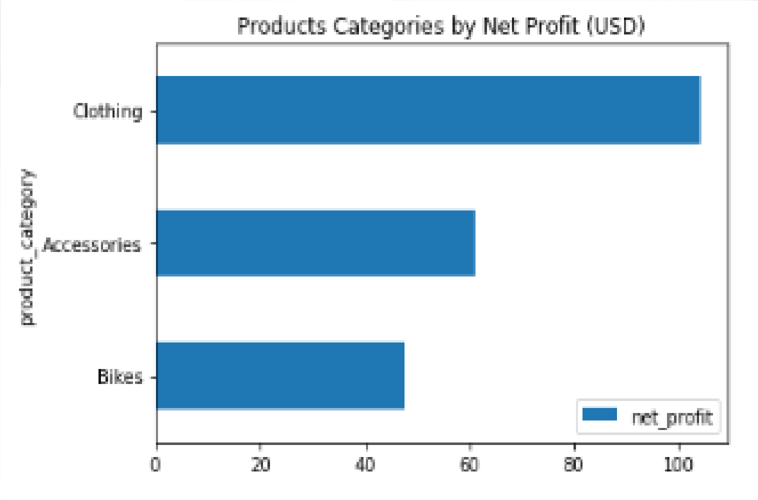

Fig4.Illustrated the products' categories and net profit.

The figure above presented the product categories and net profit for each. The clothing category had the highest net profit for the business followed by accessories, and the bike was the last.

9. Model Development

Multilinear regression was used as a method for model development. Four attributes were selected as independents variables in model design as follows: visit_month, customer_age, customer_gender, and country. Net_prifit as a response variable as shown in the formula below:

Net_profit = B1*visit_month + B2*customer_age + B3*customer_gender +B4*country + E

The data was divided into two parts training set as 0.70 and testing set as 0.30 percent.

10. Model Evaluation

Table2.Presents the model evaluation errors.

Error Measure	Value
Mean Absolute Error	15.6
Root Mean Squared Error	19.1

11. Conclusion:

We praise God and thank Him for His grace, grace and mercy. Here we are writing with our pens the last lines of this research after a great journey of effort, fatigue and staying up late.

This was an enjoyable trip worthy of fatigue and trouble, and it was a journey that elevated the mind and the mind, and it came up with important ideas for this subject. If we err, then it is from ourselves and Satan. And this is one of the greatest lessons, and it is evidence of the inferiority taking over the whole of humankind. Finally, we have advanced with ease in knowledge, and we hope that we have succeeded and obtained your satisfaction, and may God's peace and blessings be upon our master Muhammad, the illiterate prophet, the best teacher, the guide and the envoy, as a mercy to the worlds, our master Muhammad and his family and companions all.

12. Results

- Multiple linear regression model wasdeveloped to predict the customer net profit for each visit with lower error rates as illustrated in table 2 above.
- The best collections of attributes were: visit_month, customer_age, customer_gender, and country. That they were used todetermine the response variable (net_profit).
- From fig 1, it was clear that Bike Racks, Fenders, and Gloves were the top three profitable subcategories.
- According to fig 3, mountain bikes, and touring bikes were the top two subcategories that bring the highest revenue.
- As illustrated in fig 4, the clothing category had the highest net profit for the business followed by accessories.

13. Discussion

Multilinear regression is a useful method to be used for estimating the net profit for customers according to the selected features. The model hadfour features were: visit_month, customer_age, customer_gender, and country. Which determine the response variable net profit, this model should be tested over time and measure the relationship between the features to find whether additional attributes to be included or removed from the existing ones.

Customer experience has been one of the hot research areas, and many studies were conducted regarding customer next visit, revenue, purchases, responses, customer segmentation. All the studies aimed to understand customer behavior. But, the objective of this paper was to predict the customer net profit for each visit which was not covered in the literature review by previous works.

14.Future work

- More research is needed to understand customer behavior deeper.That will help in providing customers with more quality services.
- More analysis should be done especially with new technology as much data is being generated from retail transactions like big data.
- The model should be tested over time for including or excluding features.

References

(1) J. A. Lara, D. Lizcano, A. Pérez, and J. P. Valente, "A general framework for time series data mining based on event analysis: Application to the medical domains of electroencephalography and stabilometry," *J. Biomed. Inform.*, vol. 51, pp. 219–241, 2014.

(1) A. Martínez, C. Schmuck, S. Pereverzyev, C. Pirker, and M. Haltmeier, "A machine learning framework for customer purchase prediction in the non-contractual setting," *Eur. J. Oper. Res.*, 2018.

(2) G. K. Hunter and W. D. Perreault, "Making sales technology effective," *J. Mark.*, vol. 71, no. 1, pp. 16–34, 2007.

(3) J. A. Lara, D. Lizcano, A. Pérez, and J. P. Valente, "A general framework for time series data mining based on event analysis: Application to the medical domains of electroencephalography and stabilometry," *J. Biomed. Inform.*, vol. 51, pp. 219–241, 2014.

(4) J. K. Von Bischhoffshausen, M. Paatsch, M. Reuter, G. Satzger, and H. Fromm, "An Information System for Sales Team Assignments Utilizing Predictive and Prescriptive Analytics," *Proc. - 17th IEEE Conf. Bus. Informatics, CBI 2015*, vol. 1, pp. 68–76, 2015.

(5) J. Asare-Frempong and M. Jayabalan, "Predicting customer response to bank direct telemarketing campaign," *2017 Int. Conf. Eng. Technol. Technopreneurship, ICE2T 2017*, vol. 2017–Janua, pp. 1–4, 2017.

(6) P. R. Dean, D. Xue, and Y. L. Tu, "Prediction of manufacturing resource requirements from customer demands in mass-customisation production," *Int. J. Prod. Res.*, vol. 47, no. 5, pp. 1245–1268, 2009.

(7) A. De Keyser, J. Schepers, and U. Konuş, "Multichannel customer segmentation: Does the after-sales channel matter? A replication and extension," *Int. J. Res. Mark.*, vol. 32, no. 4, pp. 453–456, 2015.

(8) Q. Hu, S. Xie, J. Zhang, Q. Zhu, S. Guo, and P. S. Yu, "Hetero sales: Utilizing heterogeneous social networks to identify the next enterprise customer," *25th Int. World Wide Web Conf. WWW 2016*, pp. 41–50, 2016.

(9) F. Jaramillo and D. B. Grisaffe, "Does customer orientation impact objective sales performance? Insights from a longitudinal model in direct selling," *J. Pers. Sell. Sales Manag.*, vol. 29, no. 2, pp. 167–178, 2009.

(10) S. Kim and J. G. Lee, "A systematic framework of predicting customer revisit with in-store sensors," *Knowl. Inf. Syst.*, 2019.

(11) J. S. Leonard, D. I. Levine, and A. Joshi, "Do birds of a feather shop together? the effects on performance of employees' similarity

with one another and with customers," *J. Organ. Behav.*, vol. 25, no. 6, pp. 731–754, 2004.

(12) S. Praet and D. Martens, "Efficient Parcel Delivery by Predicting Customers' Locations*," *Decis. Sci.*, vol. 0, no. 0, pp. 1–30, 2019.

(13) R. F. Saen, "Using cluster analysis and DEA-discriminant analysis to predict group membership of new customers," *Int. J. Bus. Excell.*, vol. 6, no. 3, pp. 348–360, 2013.

(14) C. H. Schwepker and R. J. Schultz, "Influence of the ethical servant leader and ethical climate on customer value enhancing sales performance," *J. Pers. Sell. Sales Manag.*, vol. 35, no. 2, pp. 93–107, 2015.

(15) J. van Doorn, P. S. H. Leeflang, and M. Tijs, "Satisfaction as a predictor of future performance: A replication," *Int. J. Res. Mark.*, vol. 30, no. 3, pp. 314–318, 2013.

(16) T. Van Nguyen, L. Zhou, A. Y. L. Chong, B. Li, and X. Pu, "Predicting customer demand for remanufactured products: A data-mining approach," *Eur. J. Oper. Res.*, 2019.

(17) V. Vogel, H. Evanschitzky, and B. Ramaseshan, "Customer equity drivers and future sales," *J. Mark.*, vol. 72, no. 6, pp. 98–108, 2008.

(18) G. K. Uyanık and N. Güler, "A Study on Multiple Linear Regression Analysis," *Procedia - Soc. Behav. Sci.*, vol. 106, pp. 234–240, 2013.

(19) D. Cogoljević, M. Gavrilović, M. Roganović, I. Matić, and I. Piljan, "Analyzing of consumer price index influence on inflation by multiple linear regression," *Phys. A Stat. Mech. its Appl.*, vol. 505, pp. 941–944, 2018.

(20) Y. S. Kim, S. Seok, J. S. Lee, S. K. Lee, and J. G. Kim, "Optimizing anode location in impressed current cathodic protection system to minimize underwater electric field using multiple linear regression analysis and artificial neural network methods," *Eng. Anal. Bound. Elem.*, vol. 96, no. March, pp. 84–93, 2018.

(21) T. Shepel, B. Grafe, P. Hartlieb, C. Drebenstedt, and A. Malovyk, "Evaluation of cutting forces in granite treated with microwaves on the basis of multiple linear regression analysis," *Int. J. Rock Mech. Min. Sci.*, vol. 107, no. July 2017, pp. 69–74, 2018.