PREDICTING NEW CIGARETTE LAUNCH STRATEGY BASED ON SYNTHETIC CONTROL METHOD

Yu-HuaMo[1], Chao Deng[1], Fei-Jie Huang[1], Qian Tan[1*], Yuan-Kun Li[2]

[1]China Tobacco Guangxi Industrial Co., Ltd.Nanning, China
[2]PBC School of Tsinghua University,Beijing, China

ABSTRACT

In order to accurately predict the effect of new product cigarette marketing strategy.We take 18 months of cigarette sales data in city B of province A as the research sample, take new cigarette C as the research object, and use the random forest method to fix the errors and missing data. Then, we first use the mature cigarette brand's short-term historical sales and multiple labeling systems including the mature cigarette brand's historical sales data, retailer sales data, merchant circle crowd portrait data. Based on various machine learning method, we calculate the fitting weights of mature cigarettes to new cigarettes and then simulate and predict the sales trend of new cigarettes. The application effect test found the accuracy of new cigarette sales prediction based on the traditional LSTM model was only 33.31%. In comparison, the prediction accuracy of the new model we constructed can reach 94.17%. We address the limitations encountered in new cigarette sales prediction, andfill the research gap in new cigarette launch models.

KEYWORDS

Cigarette sales forecast; Synthetic control method; Machine Learning; Multiple labeling systems

1. INTRODUCTION

With the continuous development of big data-driven artificial intelligence and the financial industry, more and more scholars have started using machine learning methods to analyze and study cigarette-related predicting problems empirically. Qin Rong and Wang Jibin[1] et al. argue that the implementation of the "Internet+" action plan and the supply-side structural reform of China tobacco have a positive impact on the quality and efficiency of tobacco.Regarding the impact of machine learning in cigarette sales forecasting, Deng Chao and Liu Song[2] proposed an intelligent cigarette delivery model based on LSTM (Long Short-Term Memory) and BP (Back Propagation) neural network. Sun Jing proposed an image recognition technology-based consumer cigarette brand cultivation path algorithm.Feng Zhe and Wang Zhigang[3] proposed the use of support vector machine to solve the nonlinear and high-dimensional identification problem and build a cigarette precise placement decision model. Huang Hengbo and Chen Haiyong[4] et al. used machine learning recommendation algorithm to optimize the high-end cigarette designated placement strategy of Zhangzhou Tobacco Company. Currently, the tobacco industry has made many efforts and attempts in digital marketing, and existing methods generally focus on the use of time series to forecast the sales volume of mature cigarette products[2].

David C. Wyld et al. (Eds): ARIN, CSITA, ISPR, NBIOT, DMAP, MLCL, CONEDU -2023
pp. 33-43, 2023. CS & IT - CSCP 2023 DOI: 10.5121/csit.2023.130704

However, little literature explores new cigarette products' launch models and sales volume forecasting.

From a market perspective, there is a forecasted demand for new cigarette launches by each tobacco company. But the forecasting model for mature cigarette products cannot simply be applied to the sales forecasting of new cigarettes. There are two main reasons for this. First, mature cigarette product forecasting models must have long historical cigarette sales data. However, new cigarette products often do not have long time series data, and the short period of sample data can lead to a biased estimation of the time series model. Second, the overall trend of sales of mature cigarette products is more stable. However, due to the promotional factors and customers' newness in the early sales stage, the data of new cigarettes tend to fluctuate sharply in the early sales stage but level off in the later stage. Using time-series data to forecast the sales fluctuations of new cigarettes can easily amplify the estimation bias.

In summary, predicting the launch model of new cigarettes under the limitation of data interval is an important issue that needs to be addressed in the current industry development. In this paper, we take the new cigarette C produced in province A in 2020 as an example and use 18 months of cigarette sales data at the retailer level in city B of province A as the research sample. Our core idea is to calculate the fitting weights of mature cigarette products with new cigarettes by synthetic control method. And we use the machine learning method to calculate the sales prediction data of various mature cigarette products. Finally, the sales trend of new cigarettes is simulated based on the forecast data of mature cigarette products obtained by machine learning and the weights obtained by the synthetic control method. The accuracy of our algorithm is as high as 94.17%. We solves the current problem of the limitation of new cigarette sales prediction and fills the research gap of the new cigarette launch model.In the future, we will build a more abundant and perfect label system to improve the degree of fitting the features of new cigarette products and improve the prediction accuracy.

The arrangement of us is as follows. The first part is the introduction, the second part presents the design scheme of the prediction model based on the synthetic control method, and the third part presents the practical application effect of the prediction model. The last part is the conclusion.

2. STUDY DESIGN

This section introduces the design scheme of the forecasting model based on the synthetic control method, including basic ideas, model design, data selection, and processing, model forecasting process, and model evaluation.

2.1. The Basic Idea of the Synthetic Control Method

In the past few years, synthetic control methods have been widely used to study issues such as immigration policy, corporate political affiliation, and tax policy.The synthetic control method was initially proposed by Abadie and Gardeazabal (2003)[5], and Abadie et al. (2010)[6], to estimate how an intervention implemented at the aggregate level affects a large set of individuals (e.g., a city, region or country). The idea behind the synthetic control approach is that when the object under examination is a small number of aggregates, using the set of unaffected individuals as a control is more effective than using a single unaffected individual as a control. The synthetic control method can select the optimal linear combination of weights based on the similarity between data and is now widely used in many aspects of socioeconomic and market management. Jun Tang and Yu Gao[7] et al. used synthetic control method to evaluate the emission reduction effect of smart energy construction in smart cities. Yang Feng and Yang Wang[8] et al. used synthetic control method to explore the changing trend of China's industrial.Yujun Lian and Xin Li[9] used synthetic control method to promote the placebo test. Feiyu Chen and Qirui Chen[10] used the synthetic control method to analyze the implementation effect of waste separation-related policies in China. Lianna Yang and Yong Xu[11] used the synthetic control method to analyze the import and export trade effects of China's free trade zone. In addition, foreign literature has commonly used synthetic control methods fused with machine learning algorithms to investigate carbon market pricing, pollution emissions, and economic growth[12-14] .

2.2. Design of New Cigarette Prediction Model Based on Synthetic Control Method

We based on the synthetic control method, the new cigarettes are generated by simulating the mature cigarette products from the cigarette gauge data, calculating the weights of each mature cigarette product, and then comparing the degree of fit of the synthetic control method with the historical data to calculate the sales forecast data of the new cigarettes. The specific model design is as follows.

Assume that there are J cigarette brands. Assume that the first cigarette brand (j=1) is the target cigarette brand for which this application wants to forecast sales, and the rest (j=2, 3, ..., J) are mature cigarette brands, and the time of the data is t. For each cigarette brand j and time t, the matching variable vector is cigarette sales, cigarette brand, cigarette category, cigarette gauge, price range, and retail price of a pack. For cigarette brand j and time t, define X_{jt} to represent the above matching variable values. For the target cigarette brand j at time t, x_{1t} represents the vector of matching variables, which still includes the sales volume, cigarette brand, cigarette category, cigarette gauge, price range, and retail price of a pack of cigarettes. In order to get the fitting weights of the reference cigarette brand to the target cigarette brand, the weights obtained by the synthetic control method can be expressed by the weight vector ω_j of J×1, that is, $\omega_j=(w_2,..., w_{J+1})$, where the weights are non-negative and the sum of the weights is 1. The selection of the weights is the most important part of the synthetic control method, and the results of the research method are meaningful only when the reference cigarette brand can be a good fit to the "real area", and the equation only needs to find the minimum value of x_{1t} -$X_{jt}\omega_j$. The weights ω_j need to be chosen so that X_{jt}, which includes cigarette sales, cigarette brands, cigarette categories, cigarette sizes, price ranges, pack retail prices, and other variables, is as close to x_{1t} as possible, and the matching characteristics of the reference cigarette brands will be as close to the target cigarettes as possible after weighing. To measure this distance, a quadratic (distance between two points in a Euclidean-like space) matching equation can be used. In the process of determining the fitting weights that minimize the prediction error based on the synthetic control method, the fitting weights are called intermediate weights, and the intermediate weights are the dynamic change values in the process of determining the fitting weights.

Based on the vector of matching variables for the reference cigarette brand X_{jt}, (j=2, 3, ..., J) and the matching variable vector x_{1t} of the target cigarette, we get the matching weights ω_j of the reference cigarette brand to the target cigarette, where V is a (k*k) dimensional diagonal matrix. As shown in Equation (2).

According to the matching variable vector X_{jt} of the reference cigarette brand and the matching variable vector x_{1t} of the target cigarette brand and Equation (1), the first pre-determined formula, the prediction error τ_{1t} of the intermediate predicted sales volume, is obtained. At time t, a set of intermediate weights is selected to minimize the prediction error τ_{1t} of the synthetic control method, and the intermediate weights in this set are used as the fitting weights of each reference cigarette brand respectively.

$$\min_{\omega}(x_{1t} - X_{jt}\omega_j)'V(x_{1t} - X_{jt}\omega_j) \quad (1)$$

$$s.t.\ \omega_j \geq 0, j = 2, \ldots, J+1; \sum_{j=2}^{J+1} \omega_j = 1 \quad (2)$$

2.3. Data Selection and Processing

The four types of data used in this paper are obtained from the Tobacco Monopoly Bureau of City B in Province A, which include: the retailer's category-monthly sales dimension data, retailer's labeling dimension data, portrait labeling dimension data of the merchant circle population, and labeling dimension data of the category. Considering that the new cigarette C was launched in 2020, we chose January 2021-June 2022 for this paper.

The core data is the retailer-cigarette category-monthly sales dimension, including the monthly tracking data of 90 types for each retailer. The label "sales (units)" is the target predicted in this paper. Compared with the prediction by time series alone, we introduces a model labeling system consisting of merchant area crowd portrait, cigarette category, and retailer sales ability data to forecast cigarette sales. Compared with the time series data autocorrelation, the model label system can use the external factors that affect cigarette sales, such as shopping area population, cigarette size, and retailer characteristics, and thus better meet the unbiased and consistent data estimation.

2.3.1. Handling of Missing Data Values

In order to ensure the representativeness of the model and to eliminate the influence of extreme values within the data and the differences between different scales, the missing data are first processed.

Firstly, the data including raster data matching, and fuzzy matching of the names of the differently called gauges will be performed to spell the above four data into a data set of about 10GB, thus starting the overall data cleaning. Based on the results obtained by the synthetic control method, ten types of cigarettes were selected from the data,the details are shown in table2. The columns with more zeros and variance of less than four are removed from the training data set A to improve the calculation speed of the model.

Secondly, the size of the preliminary cleaned dataset A is 450MB. At this time, the variable names of the complete dataset A are analyzed. A variety of data with different names but synonyms and inconsistent data types are found. These data are still deleted to improve the speed of the model.

Thirdly, at this time, the complete data set A variables are filtered, and the data of type "object" is filtered out, which must be converted to int or float type to be fed into the model. For the latitude and longitude data, we can consider replacing them with the shopping area ids. Figure 1 shows the results of the visualization of shopping district ids, where the blue points are the distribution of 70% of the samples in the training set,the yellow dots are the distribution of 30% of the samples in the test set, and it can be found that the test set and the training set is more similar.

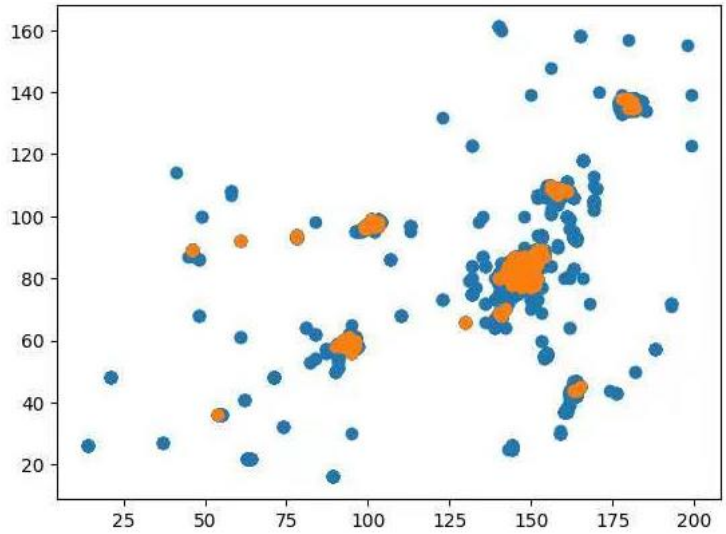


Figure 1. Business circle id visualization results

At this time, the missing values of dataset A are counted in proportion, as shown in Figure 1 below, it can be seen that "for commuting mode self-driving tgi" and "visiting 4s store brand Wuling tgi" have the largest proportion of missing values, while a total of 211 tags need to be The missing values need to be filled.

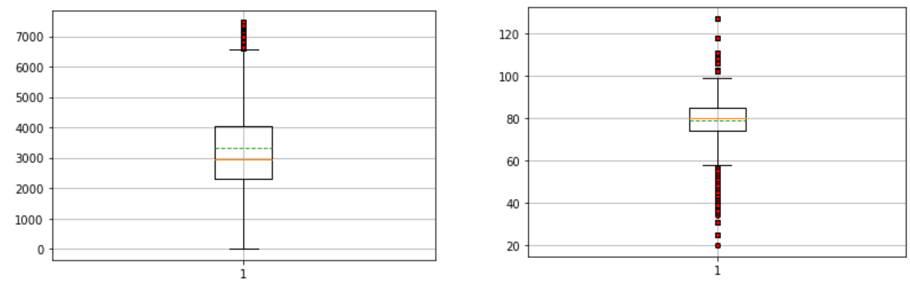
Table 1. Missing values after initial cleaning of dataset A.

	Missing Values	**%of Total**
Commuting self-driving mode tgi	23000	9.5
Visiting 4s store brand Wuling tgi	22359	8.9
Life stage retirement tgi	22245	8.9
Traveling abroad tgi	21509	8.6
...	...	...
Holiday passenger flow hour_avg9	2077	0.8
Working population	2077	0.8
Mobile passenger flow day_avg	2077	0.8
Holiday passenger flow day_avg	2077	0.8
Residential population	2077	0.8

The missing values here are considered using the random forest to fill in the missing values. Random forest is a common method in machine learning, which is to build a forest in a random way, and when a new input sample comes in, let each decision tree in the forest make a little judgment separately to see which class the sample should belong to. This method was first proposed by Breimanz in 2001[20]. We use the random forest method because there is a connection between x and y, so that x can be used to predict y. In turn, y can also predict x to some extent. When a feature x in X has a missing value, the feature is considered as target and y as a new feature. The samples without missing values are used as the training set and the samples with missing values are used as the test set, and the missing values are predicted using random forest modeling. When there are multiple features with missing values in X, we start from the feature with the least number of missing values, and fill in the other missing values with zeros. When the missing values of the feature are predicted by random forest, they are filled in the original data, and then we continue to process the next missing value as above. After the missing values are filled by random forest , and the complete cleaned data set A is obtained.

2.3.2. Handling of Data Outliers

For dataset A, we select a box plot-based outlier monitoring algorithm for monitoring. The box plot is a relatively common outlier detection method. In general, the distance between the 25% quantile Q1 and 75% quantile Q3 of all samples is taken as the length IQR of the box, and the sample values smaller than Q1-1.5 IQR or larger than Q3+1.5 IQR can be considered as outlier samples. The specific effect is shown in the following figure2.



(a) Sales outlier monitoring (b) App preference travel tgi outlier monitoring

Figure 2. Schematic diagram of outlier monitoring

2.4. Model Evaluation

In order to compare the prediction effects among models, we design the final index to evaluate the overall effectiveness of the model as the accuracy rate[3] , which means the difference between the actual and predicted sales as a ratio of the predicted sales to 1. The larger this indicator is, the better the model effect and its specific expression.

$$\text{Accuracy rate} = \left(1 - \frac{|\text{Actual Sales} - \text{Forecast Sales}|}{\text{Forecast Sales}|}\right) \times 100\% \qquad (3)$$

The synthetic control method was processed using Stata17 to calculate the weights of each mature cigarette product on the new cigarettes. The image processing, model building, and numerical computation codes are all in python language, the neural network is built using the open source Pytorch 1.10.0, and the Graphic Processing Unit (GPU) is NVIDIA GeForce RTX 3090 (NVIDIA Corporation, USA).

3. APPLICATION EFFECT TEST

This section presents the practical application effects and prediction accuracy of the prediction model based on the synthetic control method and compares it with the traditional LSTM model.

3.1. Descriptive Statistics ofthe Data

We first present descriptive statistics of the retailer-cigarette gauge-monthly sales data after the data cleaning process. Figure 3 shows the change in overall cigarette sales in City B during the sample period. From the figure, it can be seen that the overall cigarette sales have obvious seasonal fluctuations. The peak season for cigarette sales is January, September, July, August and

October, which is consistent with the change in sales volume due to seasonally influenced consumer psychological fluctuations.

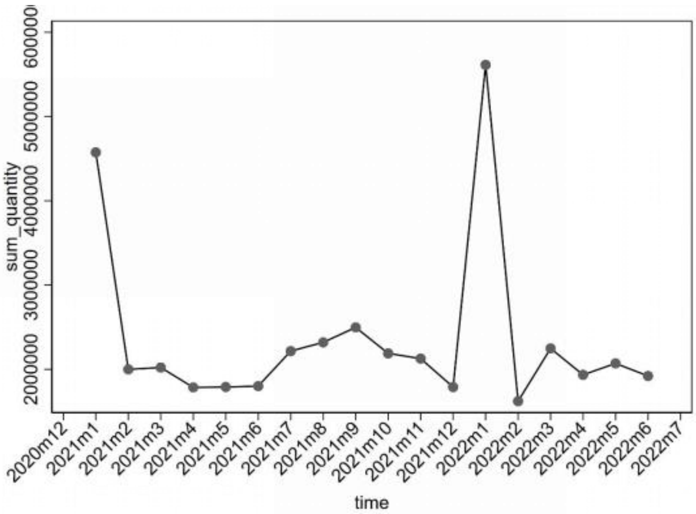


Figure 3. Changes in overall cigarette sales volume in B City

3.2. Analysis of the Prediction Results of the Traditional LSTM Model

Based on the LSTM model to model the new cigarette sales of each retailer, the LSTM model is used to learn the data for 12 months of 2021 and predict the data for January 2022-June 2022. Figure 4 shows where the blue dash indicates the real data and the yellow dash indicates the predicted results using LSTM only. Through the calculation, this paper found that the prediction accuracy of new cigarettes based on LSTM model is only 33.31%.

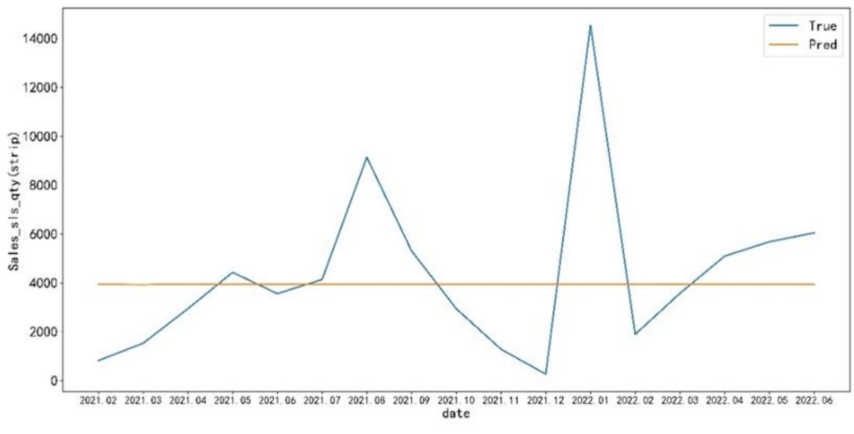


Figure 4. Prediction results based on LSTM model

3.3. Analysis of the Prediction Results of the Synthetic Control Method Model

First, we uses the synthetic control method to calculate the fitting weights of each mature cigarette product to new cigarettes. Specifically, we treat the new cigarette C as the experimental group in the data set and 89 other mature cigarette products as the control group. We selected cigarette brand, cigarette category, cigarette gauge, price range, retail price of a pack, and other cigarette gauge data as synthetic control variables. After observation, we can find that the control group with the largest weight selected by the synthetic control method is "Cordyceps sinensis (double medium)", followed by "Huangshan (small red square stamped medium)".

Table 2. Correlation coefficients between the mature product with the new product C

Cigarette Product Specifications	Correlation coefficients between the mature product with the new product
Jiaozi (wide and narrow luck)	0.002
Jiaozi (Soft Sunshine)	0.024
True Dragon (Soft Jiaozi)	0.009
True Dragon (Bama Tensei)	0.053
True Dragon (Sea Rhythm)	0.022
True Dragon (Soft Sea Rhythm)	0.136
True Dragon (HongYun)	0.101
Huangshan (Large Red Square Seal)	0.075
Huangshan (Small Red Square Seal Middle Branch)	0.188
Cordyceps sinensis (double medium branch)	0.389

Secondly, we used five machine learning methods from Weighted Ensemble, NeuralNetTorch, LightGBMLarge, RandomForestMSE, and CatBoost to compute the sales prediction data for various mature cigarette products, respectively. Table 3 shows the results run through the machine learning methods on this dataset.

Third, the predicted sales of the target cigarette brands are obtained based on the reference predicted sales and the corresponding fitting weights. Table 4 shows new cigarette sales prediction accuracy based on the fitting weights obtained from the synthetic control method model after the sales prediction data of various mature cigarette products based on each of the above machine learning methods. It can be found that the accuracy of the synthetic control method model reached 94.2% based on the Weighted Ensemble algorithm. The other machine learning models were also more than 60 percent accurate.These prediction accuracies were significantly higher than the accuracy of 33.31% for new cigarette sales based on the traditional LSTM model, proving that the new cigarette sales prediction model constructed based on the synthetic control method has a better prediction accuracy.

Table 3. List of results of each machine learning model run

Model Name	RMSE	Predicted Time(S)	Training Time (S)	Memory Usage(MB)
WeightedEnsemble	1.940	32.277	990.833	4243.000
NeuralNet	2.019	0.092	432.388	1144.000
RandomForest	2.769	0.061	1461.305	57607.000
LightGBM	2.778	0.028	58.328	27929.000
CatBoost	3.055	0.004	9.905	11223.000

Table 4. Accuracy performance of synthetic control method-based models

Algorithm Name	Prediction Accuracy
WeightedEnsemble	0.942
RandomForestMSE	0.901
NeuralNetTorch	0.891
XGBoost	0.706
LightGBM	0.691

Finally, we also compare the predicted and actual cigarette sales of new cigarettes under the synthetic control method and machine learning prediction model. Figure 6 compares predicted and actual new cigarettes, where the orange line is the predicted sales volume based on the synthetic control method, while the blue line is the actual sales volume of new cigarettes. From Figure 5, the model fits the trend and direction of new cigarettes well.

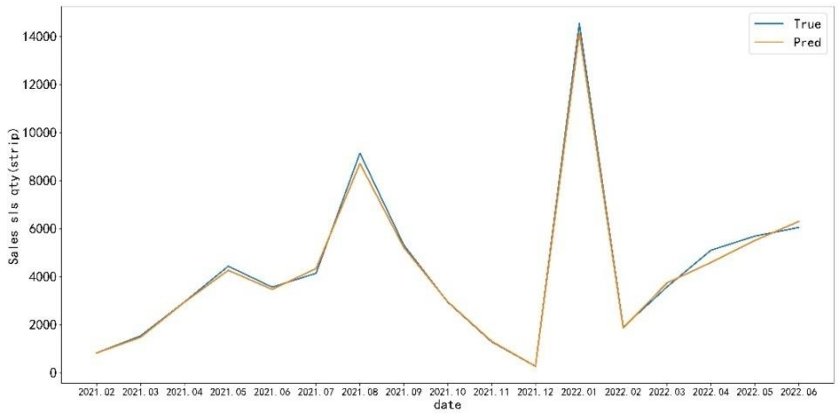


Figure 5. Comparison of predicted and actual new products in B City

4. CONCLUSIONS

To address the problem of accurate new cigarette delivery, we propose an intelligent cigarette delivery model construction method for commercial tobacco companies. The method is based on the multiple labeling systems, including historical sales data of mature cigarette brands, retailer sales data, shopping area population portrait data, and short-term historical sales of new cigarette brands, etc., combined with the fitting weights of mature cigarette product sales to new cigarette sales calculated by synthetic control method, to predict the new cigarette sales strategy. The results show that compared with the traditional LSTM model, the accuracy of the latest sales prediction model constructed based on the synthetic control method is significantly improved by 2-3 times.

Our marginal contributions are in the following three areas based on past research and the above analysis. First, we expand the literature related to the prediction of new cigarette products. We use a synthetic control method and multiple labeling systems to match and weigh new cigarette and mature cigarette product characteristics. Our prediction model enriches the theoretical research on new cigarette prediction, and fills the model gap for new cigarettes.

Second, we extend the application of existing machine learning models for sales prediction. We use various machine learning models to feed the combined data based on retailer characteristics, product gauge characteristics, and merchant area population characteristics after data cleaning and feature engineering into the model, construct the model and predict the cigarette sales of established varieties, and finally obtain the output of the machine learning model.

Third, the new cigarette prediction model obtained based on the synthetic control method meets the demand for predicting the effect of new cigarette launch strategies in the tobacco industry and has some practical significance. Our model can make customized forecasts for new cigarette products in specific categories and provide more practical guidance strategies for cigarette companies to help them make scientific decisions and launch new cigarettes scientifically.

ACKNOWLEDGEMENTS

We would like to extend our heartfelt thanks to China Tobacco Guangxi Industrial Co., Ltd for supporting our research project. Their generous funding and resources have been critical in enabling us to make significant progress and achieve our research objectives.

REFERENCES

[1] Qin Rong, Wang Jibin, Liu Jun, Wang Jue& Jiang Lengwei, (2016) "Discussion on China's Tobacco Industry Promoting "Internet Plus" Action Plan and Supply-side Structural Reform", *Market Forum,*Vol.09, pp64-69.

[2] Deng Chao, Liu Song, Wang Ludi, Gong Qiang, Gao Lin, Zuo Shaoyan, Gu Zuyi &Liang Hailing, (2021) "A method for building an intelligent cigarette delivery model based on a deep neural network", *Tobacco Science and Technology,* Vol.54, No.02, pp.78-83. DOI: 10.16135/j.issn1002-0861.2020.0349.

[3] Feng Zhe &Wang Zhigang,(2018) "Establishment of Cigarette Launch Decision Model Based on Support Vector Machine", *China Economic and Trade Guide (Middle),* Vol.32, pp103-105.

[4] Huang Hengbo, Chen Haiyong &Guo Weibin,(2020) "Research on high-end cigarette delivery strategy based on recommendation algorithm", *China New Communications,* Vol.22, No.17, pp.231-232.

[5] Abadie A &Gardeazabal J,(2003) "The economic costs of conflict: A case study of the Basque Country", *American economic review,* Vol.93, No.1, pp.113-132.

[6] Abadie A, Diamond A & Hainmueller J,(2010) "Synthetic control methods for comparative case studies: Estimating the effect of California's tobacco control program", *Journal of the American statistical Association,* Vol.105, No.490, pp.1493-505.

[7] Tang Jun, Gao Yu & Li Penglin, "Can smart cities promote the realization of dual carbon targets through smart energy construction?——Experimental evidence based on synthetic control method."

[8] Feng Yang, Wang Yang&Chang Zhongze, (2022) "Rediscussion on the Change Trend of my country's Industrial Proportion——An Empirical Analysis Based on the Synthetic Control Method", *Economic Vertical and Horizontal,*Vol.09, pp.177-86.DOI:10.16528/j.cnki .22-1054/f.202209077.

[9] Lian Yujun& Li Xin, (2022) "Improvement of Placebo Test in Synthetic ControlMethod—Statistical Inference Based on Quasi-standardized Transformation", *Statistical Research,* Vol.39, No.8, pp.14.

[10] Chen Feiyu, Chen Qirui&Li Qianwen, (2022) "Effects of Compulsory Sorting Policyof Municipal Solid Waste in China Based on Synthetic Control Method", *Chinese Population Resources and Environment,*Vol.008, pp.032.

[11] Yang Lianna&Xu Yong , (2022) "The Import and Export Trade Effect of China's Pilot Free Trade Zone——An Empirical Analysis Based on the Synthetic Control Method", *Journal of Henan University of Technology: Social Science Edition,* Vol.38, No.3, pp.9.

[12] Abrell J , Kosch M &Rausch S , (2022) "How effective is carbon pricing?—A machine learning approach to policy evaluation", *Journal ofEnvironmental Economics and Management,* Vol.112,102589.

[13] Chen Z , Kahn M E &Liu Y , et al, (2016) "The Consequences of Spatially Differentiated Water Pollution Regulation in China", Working Papers.

[14] Guo J & Zhang Z , (2019) "Does renaming promote economic development? New evidence from a city-renaming reform experiment in China",*China Economic Review,* Vol.57.

[15] Prokhorenkova L, Gusev G&Vorobev A , et al, (2017) " CatBoost: unbiased boosting with categorical features".

[16] Liaw A&Wiener M , (2002) "Classification and Regression by randomForest", *R News,* Vol.23, No.23.

[17] Qi M , (2017) " LightGBM: A Highly Efficient Gradient Boosting Decision Tree",*Neural Information Processing Systems.* Curran Associates Inc.

[18] Graves A & Jürgen Schmidhuber, (2005) "Framewise phoneme classification with bidirectional LSTM and other neural network architectures",*Neural Networks,* Vol.18, No.5-6, 602-610.

[19] Gers, Felix A &Schmidhuber, et al. , (2000) " Learning to Forget: Continual Prediction with LSTM ", *Neural Computation.*

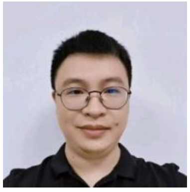
[20] Breiman. L.(2001) ."Random forests", *Machine Learning J,* Vol.45, 5-32

AUTHORS

Yuhua Mo is currently employed at China Tobacco Guangxi Industrial Co., Ltd.Herresearch interests include cigarette product recommendations, spatio-temporal data mining, and deep generative learning.

Chao Deng is currently employed at China Tobacco Guangxi Industrial Co., Ltd.research interests inmachine learning, spatio-temporal data management,and Big data visualization Technology and Application.

Feijie Huang is currently employed at China Tobacco Guangxi Industrial Co., Ltd. His research interests in Internet Entrepreneurship and Innovation System and social network data mining.



QainTan is currently employed at China Tobacco Guangxi Industrial Co., Ltd.Herresearch interests include machine learning, knowledge mapping, and software development
techniques.

Yuankun Li is currently employed at Tsinghua University. His research interests in Machine Learing and Economic Analysis.