World Academy of Science, Engineering and Technology
International Journal of Computer and Information Engineering
Vol:11, No:12, 2017

Movie Genre Preference Prediction Using Machine Learning for Customer-Based Information

Haifeng Wang, Haili Zhang

Open Science Index, Computer and Information Engineering Vol:11, No:12, 2017 publications.waset.org/10008614.pdf

***Abstract*—**Most movie recommendation systems have been developed for customers to find items of interest. This work introduces a predictive model usable by small and medium-sized enterprises (SMEs) who are in need of a data-based and analytical approach to stock proper movies for local audiences and retain more customers. We used classification models to extract features from thousands of customers' demographic, behavioral and social information to predict their movie genre preference. In the implementation, a Gaussian kernel support vector machine (SVM) classification model and a logistic regression model were established to extract features from sample data and their test error-in-sample were compared. Comparison of error-out-sample was also made under different Vapnik–Chervonenkis (VC) dimensions in the machine learning algorithm to find and prevent overfitting. Gaussian kernel SVM prediction model can correctly predict movie genre preferences in 85% of positive cases. The accuracy of the algorithm increased to 93% with a smaller VC dimension and less overfitting. These findings advance our understanding of how to use machine learning approach to predict customers' preferences with a small data set and design prediction tools for these enterprises.

***Keywords*—**Computational social science, movie preference, machine learning, SVM.

I. Introduction

THE movie industry has long relied upon various forms of data to answer questions regarding customer interest and suggested target market segments. To make informed decisions on types of movies to produce and which movie genres will be favored by specific demographic groups, data generation and analysis is necessarily rooted in all types of movie-making and streaming decisions.

A famous example of the power of data analytics in the movie industry is Netflix's algorithms used to predict the types of movies and specific movies that its individual customers are most likely to want to watch next. Netflix's online streaming service necessarily relies upon excellent customer service to maintain its customer base. This necessity requires the compilation of relevant demographic and user preference information to accurately predict movie preferences. To solve this issue, Netflix utilizes Cinematch to collect and analyze customer demographic and preference data to gain a great competitive edge over competitors like Blockbuster [1]. This type of data analytics is not only pertinent to online streaming services, however. The same type of data is needed to determine which targeted customer segments are most likely to buy or rent specific genres of movies from physical movie retail chains. Many SMEs have little or no analytical framework at hand to determine issues of marketing types of movies or whom they should be marketing. The goal of this paper is to run a predictive model usable by any SMEs in the movie rental industry in dire need of a data-based, analytical approach to answering questions regarding; 1) which customers are more likely to be interested in specific movie genres, and subsequently, 2) which market segments ought these companies target in their marketing campaigns based on the data.

A company's approach to predicting which groups of people will be interested in certain movie genres constitutes a major role in the viability of SMEs in the movie-rental market. Over the past 10 years, the number of movie-rental retail chains has drastically reduced. In 2016, only 5% of the total TV movie-rental market revenue came from physical rental stores [2]. SMEs have immense difficulty competing with established, well-known nationwide brands and need every advantage available. Simply put: SME rental businesses cannot decide which movies to promote or store without basing the decisions on predictive analysis.

Using an accurate predictive analysis model ought to be utilized by SME movie-rental businesses that do not currently employ one to reduce or prevent a loss in customer base. National consumer spending on movie rentals from physical stores has reduced from $1.22 billion in 2012 to a mere $0.49 billion in 2016 [3]. Ensuring the proper movies are stocked for the right audiences becomes even more critical when people are spending less on rentals every year.

The use of predictive analysis to ensure the right types of customers both make it into the store, as well as to help determine the correct genres of movies to stock in the store. Not only would an accurate predictive analysis retain more customers, it would reduce the amount of inventory within the store that goes unused for long periods of time. An increase in the inventory turnover and in the customer base, are two factors that directly correlate with how well a business is going. The higher the inventory turnover, the quicker movies are leaving the store to be bought or rented. A positive effect in one or both factors would result in more revenue brought in for the movie rental company.

Netflix represents an excellent example of using predictive analysis to maintain and attract new customers through targeted advertising of the right movies for the customers. The movie rental company Redbox utilizes big data to analyze trends in consumer preference. As Fig. 1 shows, Redbox has

Haifeng Wang is an Assistant Professor at the Penn State University New Kensington, 3550 7th Street Rd, New Kensington, PA 15068 (corresponding author, phone: 724-334-6726; e-mail: hzw87@psu.edu).

Haili Zhang was with Inspur Company, She is now with the Department of Information, Jinan Shandong, China, 250000 (e-mail: Zhanghali@inspur.com).

Open Science Index, Computer and Information Engineering Vol:11, No:12, 2017 publications.waset.org/10008614.pdf

utilized non-linear regression models to help propel their growth and uses predictive analysis to determine inventory of specific movies in different geographic areas [4]. This has helped spur on the massive growth that Redbox has seen. The company has used analytics to aid in its expansion from 5,000 kiosk locations in 2008 to 35,000 kiosk locations in 2014 [4]. However, relying solely upon the results of data is not a sound course of action from a strategic standpoint. Netflix often makes strategic decisions that contradict what trends data analysis has shown. For instance, data analysis showed that Netflix ought to make the subscription cancelling process more difficult for customers. Going against the data, Netflix executives decided to not place what they considered unfair burdens on the consumer to cancel their subscriptions [5].

Fig. 1 Redbox retail kiosks stock movies based on predictive analysis of trends in consumer preference

Movie rental, streaming and subscription based services are not the only types of organizations that can utilize consumer data. Educational institutions can use predictive analyses to shape some of the material taught. An emerging field in educational data mining (EDM) is being used to help educational institutions utilize data analysis to answer questions regarding certain topics, such as learning setting and popular subjects among students.

Simply put, our data cannot answer questions regarding the current mega-trends in which demographics and personal attributes contribute to an individual liking certain movie genres across the country. But, the proposed model is designed to be able to predict genre preferences within a city or a state. As [6] said, recommender users that live in South America often dislike Hollywood drama movies. Hence, the movie recommender users in different states have different movie genre preferences and recommendations for users in other parts of the world have to be refined. The reason why we did not use the 10 million dataset from MovieLens is that this dataset does not include demographic information. Each user is represented by an ID and no other personal information is provided [6].

Our dataset consists of one sample with 1100 observations regarding movie genres. Preferences for 11 types of movies, 22 personal attribute variables and five demographic variables. The question our data can answer, however, is how a small company can utilize our predictive analysis model to determine which customers of theirs will be drawn to a specific type of movie. Using survey data, we will display the many relationships existing within current company data that exists between customer preferences and demographic information to display trends in age, gender and other factors in personal movie preferences that can be used to accurately market specific movies to customers. In general, the user information such as gender, location, or preference is effectively used in movie recommendation systems [7]-[10]. In this paper, we will examine the characteristics of survey respondents who like comedy movies.

Our survey recorded individual values for their propensity to like certain types of movies or how strongly the person feels a certain characteristic applies to himself or herself. The values range from one to five, one being the lowest and five being the highest. The median age for men and women who enjoy comedy movies is 20 years old. Out of the 1008 responses for comedy movies, 507 of them are considered positive responses (a value of five) and 501 of them are considered negative responses (a value of zero to four). Breaking down the survey respondents by demographic, women who have obtained their bachelor's degrees tend to like comedy movies the most; with an average score of 4.5. This average score holds true for whether the respondents live in either of our two living categories (village or city). For the men, those who live in cities who graduated high school (but have not graduated from college) have an average comedy movie score of 4.58. For men who live in villages, the group with the highest average comedy score is college graduates with an average of 4.21; significantly less than the average response from women in the same demographic categories.

Our total data set shows some clusters in terms of education level and age. The age range for respondents in the survey is 15-29 years, however, most respondents are in the 18 years to 22 years-old age range, while a significant portion of all respondents are either current secondary students or have graduated from secondary school but have not yet graduated college; this could be they are either currently attending college, or simply chose not to attend. These trends in education level and age have the propensity to favor the movie genre we are focusing on here, comedy. We do not believe that these general trends will inhibit our predictive model in any significant way, but these trends supply baseline assumptions that the predictive model can then be run against.

In this paper, we propose an improved machine learning approach to predict movie genre preferences based on demographic information. We investigate the prediction accuracy of two machine learning algorithms including logistic regression and Gaussian Kernel SVM. The paper is structured as follows. In Section II, we describe the related methods used in movie recommendation systems. In Section III, we present the proposed predictive method. In Section IV, we evaluate experimental results and discuss prediction performance. In Section V, we leave the reader with concluding thoughts and future work.

II. Related Work

Various recommender systems have been developed to guide customers to find items that might be of interest to them. And the recommendation performance has being improved

Open Science Index, Computer and Information Engineering Vol:11, No:12, 2017 publications.waset.org/10008614.pdf

recently by different approaches [11], [12]. Researchers usually categorize recommender systems into collaborative filtering and content-based filtering systems [11], [13], [14]. This section provides a brief review of both filtering methods and known issues associated with the approaches.

A. Collaborative Filtering Approach

The collaborative filtering approach recommends the items of interest to a particular user based on the similarity to past ratings. Based on customers' preferences, a collaborative filter calculates the correlation coefficient between the customers being served with other customers. This is called Pearson correlation coefficient [9], [15], [16]. If the coefficient is near +1, this means that the two customers have similar preferences. Secondly, the approach will select neighbors who have a high coefficient for the customer. Finally, the collaborative filtering method predicts the customer's preference for a specific movie genre based on neighbors' ratings.

B. Known Problems of Collaborative Filtering Approach

The two major problems are sparsity and cold-start. The sparsity problem occurs when there are not enough customer information and ratings available. If we collect survey results only from small amount of users, the accuracy of the recommendation from trained recommendation system will be lower than the accuracy obtained based on a large number of samples [9], [10], [17]-[19]. Actually, we found the more features we used in machine learning algorithm, the more samples we need to prevent overfitting. In addition, the cold-start is the other problem to collaborative filtering. The problem happens when new customers or movies do not have enough information or rating in the recommendation system [20]-[22]. Although existing systems could become unreliable because of the cold-start problem, the recommendation model designed for SMEs is affected less than those models designed for individual customers by the cold-start issue. The model for SMEs predicts genre preferences for a group of audiences rather than one audience, so a new customer who has no information recorded will not affect the prediction for the general genre preferences of targeted customer base. However, the model for SMEs in quickly growing cities should be updated continually over time to prevent degrading as the numbers of customers increases significantly.

C. Content-Based Filtering Approach

This algorithm is based on descriptions of items and user preferences to recommend products to customers. The approach compares user's preferences with new items' representations and matches user preferences with item attributes. There are some machine learning technologies that have been applied to the content-based filtering approach, such as naïve Bayes [23], [24].

D. Drawback of Content-Based Approach

Content-based filtering algorithm has some drawbacks. First, this algorithm relies on appropriate information for categorization. It cannot generate reliable suggestions without such information. That means this content based filtering method cannot provide suitable recommendation results if the analyzed content does not contain sufficient information for classifying items. Sometimes, domain knowledge and an ontology are required to determine attributes in recommendation [25]. In addition, when enormous amounts of attributes information is calculated by matrix-based approaches, the scalability and sparsity problem may occur [26].

This work here can be viewed as a confluence and continuation of the above-mentioned works. We draw some key points from a customer-oriented recommender system and present new contributions. In particular, we extend the previous application by helping SMEs utilize data analysis to answer questions regarding customer preferences and marketing segment of the movie rental industry. The main contributions of this work are the following:

- A machine learning based collaborative filtering recommender system is presented for SMEs. The proposed approach employed two classification models to fully implement accurate movie genre prediction in quantitative and qualitative aspects.
- An analysis is performed to study how to process the dataset and select representative features to prevent the overfitting issue and improve prediction performance.
- Some suggestions are provided on how to choose machine learning classification algorithms based on samples and features.

III. Methods

In this section, we describe the methodology behind experiments that were performed.

A. Handle Diverse and Big Data

The sample shown in Table I was composed of 38 hypothetical decision-making dimensions (11 for different preference to movie genres; 27 for demographic information) with a rating ordinal scale from 1 to 5. The scale was set as below: 1-strongly bored, 2-bored, 3-whatever, 4-acceptable, 5-strongly interested. For the sake of simplicity, we only regard the scale of 5 as positive recommendation and other scales are negative ones. A sample size of 1100 respondents with quota characteristics enables the research study to be generalized to a young population.

TABLE I
RESEARCH DATA SAMPLING METHODOLOGY

Type of sampling data	Quantitative representative research
Population age	15-29 years
Population education level	secondary students to college graduates
Size	1100 respondents
Quota characteristics:	Demographic information and movie genre preference
Method	Omnibus research: personal inquiring

B. Prune Data

Before analyzing the sample of 1100 respondents, we firstly cleaned missing values in personal inquiring by ignoring incomplete observations. Data were considered sparse when

Open Science Index, Computer and Information Engineering Vol:11, No:12, 2017 publications.waset.org/10008614.pdf

the expected values in a dataset were missing. Variables with missing information represented by "N/A" or left blank completely were deleted from the data set to avoid skewing the results. This initial round of cleaning provided 1008 complete responses. Data were cleaned in R Studio.

C. Balanced Data

Out of the 1008 responses for comedy movies, 507 of them were considered positive responses (a value of five) and 501 of them were considered negative responses (a value of zero to four).

D. Investigate and Select Features

We identified predictors that separated classes well by plotting different pairs of predictors on scatter plots. The plot helped investigate classes' separation to include or exclude predictors. For example, the feature "PC" and "Finance" did not separate movie genre preferences into two classes, as in Fig. 2, and thus, both of them should be excluded from the useful features. Based on other scatter plots, the predictors in Table I were excluded.

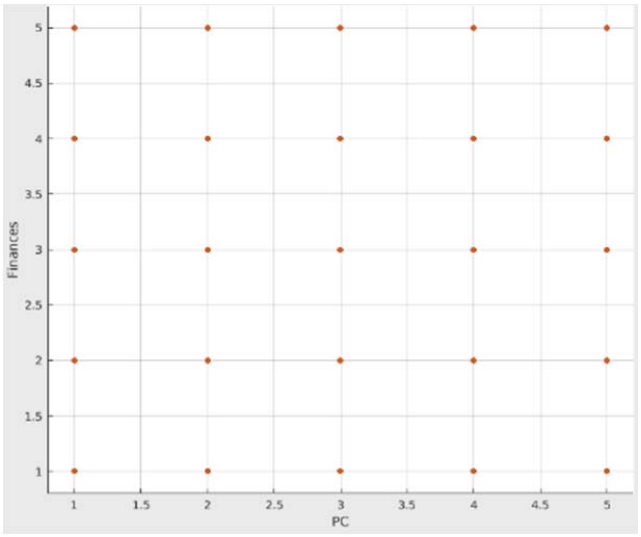


Fig. 2 This is scatter plot for "Finances" and "PC" that have positive prediction in all cases marked by red points

TABLE II
THE PREDICTORS THAT WERE EXCLUDED BECAUSE THEY CANNOT SEPARATE THE COMEDY GENRE PREFERENCE INTO TWO CLASSES

Excluded Predictors in sample dataset					
PC	CHILDREN	INTERNET USAGE	FINANCES	ONLY CHILD	VILLAGE TOWN

E. 5-Fold Cross Validation

Preference for comedy was used as a response variable; 31 predictors were used as independent variables. To prevent overfitting, one way was to not use the entire data set when training the classifier. Part of the data was removed before training begins. When the training process was finished, the removed data can be used to test the performance of the learned model on "unknown" data. The whole data set was divided into five subsets, the so-called 5-fold cross validation.

The training samples were randomly partitioned to five equal parts. We used four parts for training and the left one part for validation. So each time, one of the five subsets was used as the test set and the other four subsets were put together to form a training set. Then the average error across all five trials was computed. The advantage of the 5-fold cross validation was that it did not matter how the data sets were divided. Each observation set was used as a test set once, and used in the training set four times. The variance of the trained prediction model was reduced. The average of validation errors is called the cross validation error. We used the cross validation error in the model selection process.

F. Least Absolute Shrinkage and Selection Operator

Least absolute shrinkage and selection operator (LASSO) were used to select and regularize variables in the training.

LASSO solves the problem:

$$min_{\beta_{0,\beta}} \left(\frac{1}{2N}\sum_{i=1}^{N}(y_i - \beta_0 - x_i^T\beta\right)^2 + \lambda\sum_{j=1}^{p}|\beta_j| \qquad (1)$$

where N is the number of observations; y_i is the response at observation I; x_i is the data, a vector of p values at observation I; λ is a nonnegative regularization parameter; the parameters β_0 and β are a scalar and a vector of length p, respectively. The algorithm that LASSO used is based on the Alternating Direction Method of Multipliers (ADMM) [27]. Due to the space limitation, no further details of algorithm are presented here. The interested reader is referred to [28], [29] for a more complete description.

The LASSO algorithm was programmed to input a matrix of 800*27 dimensional predictors with redundant variables. The algorithm returned the coefficient vectors for "Happiness in life" and "Internet Usage" were zero. It meant that lasso identified the two redundant predictors in the samples, and so, we removed the two predictors in the training. Experimental evaluations have shown that using LASSO improves the prediction accuracy of preferences [30].

G. Logistics Regression Model

A logistic regression classifier was trained to classify movie preference (very interested or not) using 800 observations from a training data set. As a statistical method, Logistic regression analyzed a dataset in which there were 25 independent variables that determined an outcome. Logistic regression outcome is a dichotomous dependent variable including only two possible outcomes as 1 (TRUE, interested, etc.) or 0 (FALSE, not interested, etc.).

The objective of logistic regression algorithm was to find the best fitting model to describe the relationship between the dependent variable (movie genre preference) and a set of independent predictor variables (customers' information). Logistic regression generated the coefficients to predict a logit transformation of the probability of presence of the preference for comedy. Due to the space constrains, no further details of logistic regression are presented here. The interested reader is referred to [31], [32] for a more complete description.

Open Science Index, Computer and Information Engineering Vol:11, No:12, 2017 publications.waset.org/10008614.pdf

H.SVM Model with Gaussian Kernels

The SVM classifier obtained by solving the convex Lagrange dual of the primal max-margin SVM formulation is as:

$$f(x) = \sum_{i=1}^{N} a_i y_i K(x, x_i) + b \quad (2)$$

where N is the number of support vectors. Instead of imagining the original features of each data point, we considered a transformation from x_n space to a new feature space z_n *which was equal to* $\emptyset(x_n)$.

The data point had 25 features, one for each support vector. The value of the n^{th} feature was equal to the value of the kernel between the n^{th} support vector and the data point being classified. In this space, the original SVM classifier was just like any other linear discriminant.

Note that after the transformation, the original features of the data point were irrelevant. It was represented only in terms of its dot products with support vectors (which are basically special data points chosen by the SVM optimization algorithm). The Gaussian kernel has another name called Radial Basis Function (RBF) kernel is given as:

$$K(x, x') = \exp(-\gamma \|x - x'\|\text{^}2), \ \gamma > 0 \quad (3)$$

The linear combination of Gaussians is centered at support vectors.

$$g_{svm}(x) = sign(\sum_{sv} a_n y_n K(x, x_n) + b) \quad (4)$$

$$b = y_n - w^T z_n \ if \ a_n > 0, y_n \ is \ given \ by \ samples \quad (5)$$

So, the final hypothesis from the Gaussian SVM algorithm used Gaussian kernel and found the coefficients in the linear combination, a_n and the Gaussian function on those support vectors. The bias term b was a feasible intercept based on constraint conditions.

The SVM classifier with the Gaussian kernel is simply a weighted linear combination of the kernel function computed between a particular data point x and each of the support vectors x_n. The role of a support vector in the classification of a data point is tempered with a_n.

I. Training Models

To programmatically train a classifier, we used the flow chart in Fig. 3 to automate train the logistic regression model and Gaussian kernel SVM model with the training data. We input a table containing the predictor and response and got output for the trained Classifier and accuracy.

IV. Evaluation

In this section, we describe the results of our experiments with the machine learning models. The goal of our evaluation was to determine if the system was able to predict genres which customers mostly like. Across the paper, the accuracy is expressed using the area under receiver operating characteristic curve (AUC) coefficient and Receiver Operating Characteristic (ROC) curves. AUC is widely used in social sciences. If the AUC=0.5, the classifier is correct at half of the time. If the AUC=1, the classifier is always correct [33].

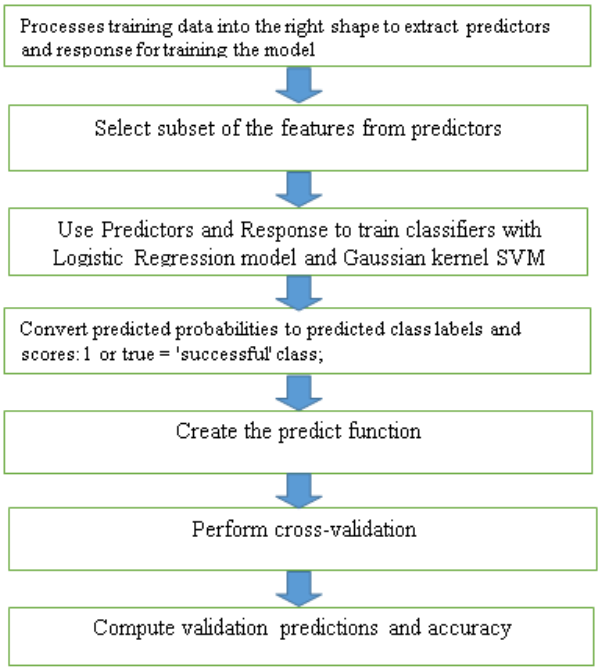


Fig. 3 Classifier training algorithm flowchart

A. Test Resources

In order to demonstrate the robustness of the prediction obtained using two proposed classifiers, 208 test samples from the same survey were used to validate our classification algorithms, so as to provide a benchmark to compare prediction accuracy between two classifiers.

B. Test Results and Discussion

We classify each response in four classes as below:

- True Positive (TP): the system suggests that customers like comedy very much, and the customers rated comedy with a five.
- True Negative (TN): the system suggests that customers will dislike comedy, and the customers rated comedy with a four or lower value.
- False Positive (FP): the system suggests that customers like comedy very much, and the customers rated comedy with a four or lower value.
- False Negative (FN): the system suggests that customers will dislike comedy, and the customers rated comedy with a five.

In order to evaluate different aspects of classifiers, we have used Information Retrieval performance metrics. In the movie genre recommendation scenario, precision is computed as the number of comedies that the system correctly predicts as the genre that the customer will like to watch (TP) divided by the total number of the comedy genre recommended positively. The general formula for Precision is:

$$Precision = tp/(tp + fp) \quad (6)$$

Open Science Index, Computer and Information Engineering Vol:11, No:12, 2017 publications.waset.org/10008614.pdf

The Recall is defined as the number of retrieved relevant resources divided by the number of relevant resources. In the genre recommendation system, it can be computed by dividing the number of genre correctly recommended by the total number of genres that are worth recommending. The formula for Recall is as follows:

$$Recall = tp/(tp + fn) \quad (7)$$

Accuracy stands for the fraction of resources predicted as the positive or negative for which the prediction was correct [6]. We use the following formula for accuracy:

$$Accuracy = (tp + tn)/(tp + tn + fp + fn) \quad (8)$$

Certainly, different applications have different precedence to precision and recall. In our application, recall was frequently regarded as more important than precision, as it was acceptable to increase the amount of false positives (FP). Some customers may also like to rent or buy comedy movies for entertainment, even though their responses on the scale of 4 (acceptable) were regarded as negative prediction in our algorithm. The test performance summary without movie genre preference predictors is listed in Table III.

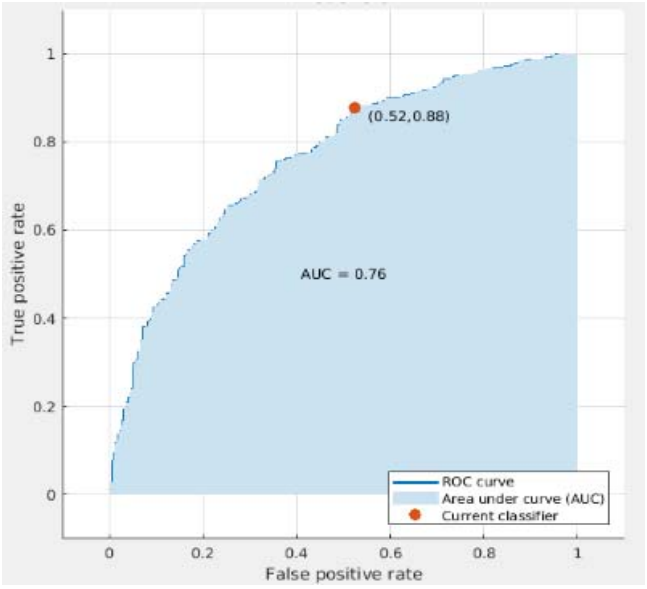
TABLE III
A SUMMARY OF TEST PERFORMANCE WITHOUT MOVIE GENRE PREFERENCE PREDICTORS

Algorithms	TP	TN	FP	FN	Precision	Recall	Accuracy
Medium Gaussian SVM	0.53	0.393	0.046	0.024	0.92	0.96	0.93
Logistic Regression	0.5	0.393	0.036	0.06	0.93	0.89	0.89

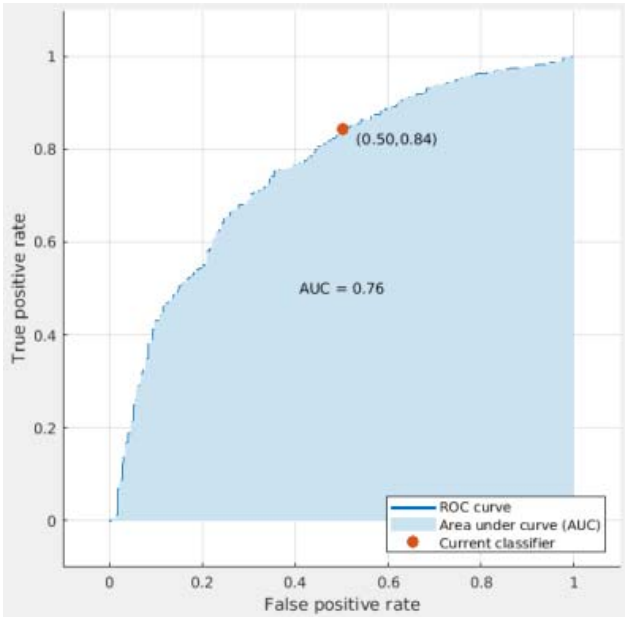
So, our predictive model enable SMEs to answer questions regarding; 1) which customers are more likely to be interested in specific movie genres, and subsequently, 2) which market segments ought these companies target in their marketing campaigns based on the data. In sum, we focus the analysis of the performance by prioritizing recall over precision and we focus the analysis on positive recommendations (in much favor of certain genre) rather than on negative ones.

In comparing the results in Fig. 4 obtained from the Gaussian SVM and logistic regression classifiers, it was fairly obvious that Gaussian SVM performed better compared to logistic regression classifier in terms of positive class irrespective of error-in-sample and error-out-sample. Under the 800 training experiments, maximum positive prediction accuracy of 88% and 84% were obtained using Gaussian kernel SVM classifier and logistic regression, respectively.

In a comparison of Tables IV and V, the movie genre preference data did not improve the accuracy as we expected but led to a surprising drop in prediction accuracy. After we discarded the movie preference data in the training and test, the data dimension decreased. In 800 new training experiments without these preference data, maximum positive prediction accuracy of 83% and 66% were obtained using Gaussian kernel SVM classifier and logistic regression, respectively. Under 208 new tests, the maximum positive prediction accuracy of 93% and 90% were obtained using Gaussian kernel SVM classifier and logistic regression, respectively.



(a)



(b)

Fig. 4 (a) Medium Gaussian SVM Positive class prediction results; (b) logistic regression Positive class prediction results

Thus, when the data dimension became larger and the VC-dimension became larger, the error-in-sample was lower but the error-out-sample was higher, and as such, bad generalization and overfitting happened. Moreover, random noise in the collected data also affected overfitting. Besides noise information in the training data, the target complexity in the training data also acted like noises; thus, that larger VC-

dimension needs more training data to avoid overfitting. To prevent overfitting when machine learning used movie genre preference information to find the target function in the hypothesis set, more data or observations and less noise in the training data were necessary. Otherwise, discarding genre information in the training was a better choice.

TABLE IV
POSITIVE PREDICTION ERROR IN TRAINING AND TESTS WHEN DATA INCLUDED MOVIE GENRE PREFERENCE SAMPLING

Error	Machine learning Algorithm	
	Medium Gaussian SVM	Logistic Regression
Error-out-sample	0.1429	0.2262
Error-in-sample	0.12	0.16

TABLE V
POSITIVE PREDICTION ERROR IN TRAINING AND TESTS WHEN DATA DID NOT INCLUDE MOVIE GENRE PREFERENCE SAMPLING

Error	Machine learning Algorithm	
	Medium Gaussian SVM	Logistic Regression
Error-out-sample	0.0714	0.107
Error-in-sample	0.17	0.34

From studies that we performed, we recommend to use Gaussian kernel SVM rather than logistic regression in machine learning when the number of features is small; for example, the feature number is less than 1000 and the training samples is intermediate, for example, the number of samples is between 100 and 10000.

V. CONCLUSION

From the studies that were performed, we can conclude that our application of machine learning techniques for movie genre prediction is quite successful. The geographic factors contain more information about movie preference prediction than that we can perceive by sex classification as usual. The experiments showed that the listed geographic information can be used to accurately predict customer movie genre preferences, for example, comedy movies. As the paper shows, features which were chosen in machine learning had an impact on the prediction accuracy; however, some features were redundant predictors in samples and need be found and removed by algorithms. The experiments also compared the prediction results between small and large VC-dimensions and showed that the accuracy was negatively correlated with the VC-dimensions when the amount of sample data was not sufficient. The lower accuracy of high VC-dimension classifier confirmed that more predictors of the information can result in overfitting and negatively influence prediction accuracy when the amount of training data was not sufficient. The findings showed that the Gaussian kernel SVM algorithm and proper predictor numbers significantly enhanced the prediction power of the user preference for movie genre. Therefore, the proposed machine learning algorithm allows us to overcome the shortcomings of a traditional massive data-based recommender system and could be enhanced by more local customer samples and be used particularly by local SMEs in the movie rental industry.

REFERENCES

[1] John, G.: 'Netflix Case Study: David Becomes Goliath', in Editor: 'Book Netflix Case Study: David Becomes Goliath' (2008, September 13, edn.).
[2] Group, C.C.: 'Distribution of movie and TV rental market revenue in the United States from 2012 to 2016', in Editor: 'Book Distribution of movie and TV rental market revenue in the United States from 2012 to 2016' (2017, edn.).
[3] Group, D.E.: 'Consumer spending on home entertainment rentals in the United States from 2012 to 2016, by type (in billion U.S. dollars).', 'Book Consumer spending on home entertainment rentals in the United States from 2012 to 2016, by type (in billion U.S. dollars).' (2017, edn.).
[4] Booker, E.: 'Predictive Analytics at Work at Ebay, Redbox.', in Editor: 'Book Predictive Analytics at Work at Ebay, Redbox.' (2014, edn.).
[5] KNOWLEDGE@WHARTON: 'How Data Analytics Is Shaping What You Watch': 'Book How Data Analytics Is Shaping What You Watch' (2015, edn.).
[6] Briguez, C. E., Budán, M. C. D., Deagustini, C. A. D., Maguitman, A. G., Capobianco, M., and Simari, G. R.: 'Argument-based mixed recommenders and their application to movie suggestion', Expert Systems with Applications, 2014, 41, (14), pp. 6467-6482.
[7] Choi, S.-M., Ko, S.-K., and Han, Y.-S.: 'A movie recommendation algorithm based on genre correlations', Expert Systems with Applications, 2012, 39, (9), pp. 8079-8085.
[8] Bell, R.M., and Koren, Y.: 'Lessons from the Netflix prize challenge', SIGKDD Explor. Newsl., 2007, 9, (2), pp. 75-79.
[9] Billsus, D., and Pazzani, M. J.: 'Learning Collaborative Information Filters'. Proc. Proceedings of the Fifteenth International Conference on Machine Learning1998.
[10] Sarwar, B., Karypis, G., Konstan, J., and Riedl, J.: 'Analysis of recommendation algorithms for e-commerce'. Proc. Proceedings of the 2nd ACM conference on Electronic commerce, Minneapolis, Minnesota, USA2000.
[11] Adomavicius, G., and Tuzhilin, A.: 'Toward the next generation of recommender systems: a survey of the state-of-the-art and possible extensions', IEEE Transactions on Knowledge and Data Engineering, 2005, 17, (6), pp. 734-749.
[12] Park, D. H., Kim, H. K., Choi, I. Y., and Kim, J. K.: 'A literature review and classification of recommender systems research', Expert Systems with Applications, 2012, 39, (11), pp. 10059-10072.
[13] Lops, P., De Gemmis, M., and Semeraro, G.: 'Content-based recommender systems: State of the art and trends': 'Recommender systems handbook' (Springer, 2011), pp. 73-105.
[14] Herlocker, J. L., Konstan, J. A., Borchers, A., and Riedl, J.: 'An algorithmic framework for performing collaborative filtering', in Editor (Ed.)^(Eds.): 'Book An algorithmic framework for performing collaborative filtering' (ACM, 1999, edn.), pp. 230-237.
[15] Sarwar, B., Karypis, G., Konstan, J., and Riedl, J.: 'Item-based collaborative filtering recommendation algorithms'. Proc. Proceedings of the 10th international conference on World Wide Web, Hong Kong, Hong Kong2001 pp. Pages.
[16] Huang, Z., Chen, H., and Zeng, D.: 'Applying associative retrieval techniques to alleviate the sparsity problem in collaborative filtering', ACM Trans. Inf. Syst., 2004, 22, (1), pp. 116-142.
[17] Popescul, A., Ungar, L. H., Pennock, D. M., and Lawrence, S.: 'Probabilistic Models for Unified Collaborative and Content-Based Recommendation in Sparse-Data Environments'. Proc. Proceedings of the 17th Conference in Uncertainty in Artificial Intelligence2001 pp. Pages
[18] Wilson, D.C., Smyth, B., and Sullivan, D.O.: 'Sparsity Reduction in Collaborative Recommendation: A Case-Based Approach', International Journal of Pattern Recognition and Artificial Intelligence, 2003, 17, (05), pp. 863-884.
[19] Ishikawa, M., Geczy, P., Izumi, N., Morita, T., and Yamaguchi, T.: 'Information Diffusion Approach to Cold-Start Problem'. Proc. Proceedings of the 2007 IEEE/WIC/ACM International Conferences on Web Intelligence and Intelligent Agent Technology - Workshops2007.
[20] Schein, A. I., Popescul, A., Ungar, L. H., and Pennock, D. M.: 'Methods and metrics for cold-start recommendations'. Proc. Proceedings of the 25th annual international ACM SIGIR conference on Research and development in information retrieval, Tampere, Finland 2002.
[21] Tang, T., and McCalla, G.: 'Utilizing Artificial Learners to Help Overcome the Cold-Start Problem in a Pedagogically-Oriented Paper Recommendation System', in De Bra, P.M.E., and Nejdl, W. (Eds.):

Open Science Index, Computer and Information Engineering Vol:11, No:12, 2017 publications.waset.org/10008614.pdf

'Adaptive Hypermedia and Adaptive Web-Based Systems: Third International Conference, AH 2004, Eindhoven, The Netherlands, August 23-26, 2004. Proceedings' (Springer Berlin Heidelberg, 2004), pp. 245-254.
[22] Lew, M. S., Sebe, N., Djeraba, C., and Jain, R.: 'Content-based multimedia information retrieval: State of the art and challenges', ACM Trans. Multimedia Comput. Commun. Appl., 2006, 2, (1), pp. 1-19.
[23] Li, Y., Lu, L., and Xuefeng, L.: 'A hybrid collaborative filtering method for multiple-interests and multiple-content recommendation in E-Commerce', Expert Systems with Applications, 2005, 28, (1), pp. 67-77.
[24] Son, J., and Kim, S. B.: 'Content-based filtering for recommendation systems using multiattribute networks', Expert Syst. Appl., 2017, 89, (C), pp. 404-412.
[25] Noia, T. D., Mirizzi, R., Ostuni, V. C., Romito, D., and Zanker, M.: 'Linked open data to support content-based recommender systems'. Proc. Proceedings of the 8th International Conference on Semantic Systems, Graz, Austria 2012.
[26] Boyd, S., Parikh, N., Chu, E., Peleato, B., and Eckstein, J.: 'Distributed Optimization and Statistical Learning via the Alternating Direction Method of Multipliers', Found. Trends Mach. Learn., 2011, 3, (1), pp. 1-122.
[27] Tibshirani, R.: 'Regression shrinkage and selection via the lasso: a retrospective', Journal of the Royal Statistical Society Series B, 2011, 73, (3), pp. 273-282.
[28] Zou, H., and Hastie, T.: 'Regularization and Variable Selection via the Elastic Net', Journal of the Royal Statistical Society. Series B (Statistical Methodology), 2005, 67, (2), pp. 301-320.
[29] Pereira, A .L. V., and Hruschka, E. R.: 'Simultaneous co-clustering and learning to address the cold start problem in recommender systems', Know.-Based Syst., 2015, 82, (C), pp. 11-19.
[30] Dreiseitl, S., and Ohno-Machado, L.: 'Logistic regression and artificial neural network classification models: a methodology review', Journal of Biomedical Informatics, 2002, 35, (5), pp. 352-359.
[31] Owen, A. B.: 'Infinitely Imbalanced Logistic Regression', J. Mach. Learn. Res., 2007, 8, pp. 761-773.
[32] Wang, Y., & Kosinski, M: 'Deep Neural Networks Are More Accurate Than Humans at Detecting Sexual Orientation From Facial Images', Journal of Personality and Social Psychology, 2017, (in press).
[33] Wang, Y., & Kosinski, M: 'Deep Neural Networks Are More Accurate Than Humans at Detecting Sexual Orientation From Facial Images', Journal of Personality and Social Psychology, 2017, (in press).

Open Science Index, Computer and Information Engineering Vol:11, No:12, 2017 publications.waset.org/10008614.pdf