Machine Learning-Based Steam Platform Game's Popularity Analysis

Wenxiang Hu[1], Yiming Wang[2] and Ruoming Xia[3]
[1]*Bachelor of information Technology, The University of Newcastle, 039594, Singapore*
[2]*College of Electronic Science and Engineering, Jilin University, Changchun, 130015, China*
[3]*College of Blockchain Industry, Chengdu University of Information Technology, Chengdu, 610225, China*

Keywords: Machine Learning, Steam Games, Data Analytics, Model Prediction, Good Reviews.

Abstract: This study delves into the application of machine learning techniques to analyze game popularity on the Steam platform. Utilizing a diverse array of algorithms such as logistic regression, Support Vector Machine (SVM)，decision tree, Gradient Boosting (XGBoost)，Light Gradient Boosting Machine (LightGBM or LGBM), Deep Neural Networks (DNNs), and Convolutional Neural Networks (CNN), the research focuses on predicting game popularity through a thorough analysis of the Steam game dataset. The report meticulously outlines the stages of data preparation, including data cleaning and feature engineering, followed by the construction of various predictive models and their subsequent performance evaluation. Notably, the LGBM demonstrated a marked advantage, boasting an accuracy of 88.17% and an AUC of 80.36%. This investigation into game popularity on Steam not only aids game developers and companies in strategic planning and risk mitigation but also provides valuable insights for player community administrators to enhance community management. The comprehensive approach underscores the significant potential of machine learning in interpreting market trends and player preferences within the gaming industry.

1 INTRODUCTION

With the vigorous development of the digital entertainment industry, electronic games have become an indispensable entertainment element around the world. Further, the financial success of gaming firms is determined by the popularity of their products. Valve Corporation is the owner of the digital distribution platform for video games, Steam. It is currently the most popular platform for playing video games on computers. While many games remain unpopular, some instantly draw massive numbers of players. A game's favorable rating on Steam might be a useful indicator of its level of popularity. In practice, it is challenging to pinpoint exactly which aspects of a game's positive rating shift, and the favorable rating is heavily influenced by the player's perception and assessment of the game's numerous elements (Predicting the Popularity of Games on Steam, 2021).

Machine learning methodologies are broadly categorized into supervised and unsupervised learning, differentiated primarily by whether the data utilized is human-labeled. Supervised learning employs labeled data as a definitive learning target, often achieving effective learning outcomes. However, the procurement of labeled data can be costly. Conversely, unsupervised learning, akin to autonomous or self-directed learning, leverages a broader data spectrum. This approach may uncover more patterns inherent in the data, potentially surpassing the insights from manually labeled patterns, though it tends to have lower learning efficiency (Palma-Ruiz et al. 2022). A commonality between these two approaches is their reliance on constructing mathematical models for optimization problems. Typically, these problems do not have perfect solutions, reflecting the complexity and challenges inherent in machine learning tasks. Machine learning methods learn by having computers process large data sets to analyze the large number of fixed structures and patterns contained in the data set that can be analytically summarized. This gives the computer the ability to process new sample data sets and make predictions about them. In practice, machine learning methods have been more popularly used, with deep learning methods being more prevalent, such as Convolutional Neural Networks (CNN) with recurrent neural networks (RNN). For example, Cheng et al. in 2016, analyzed and

Hu, W., Wang, Y. and Xia, R.
Machine Learning-Based Steam Platform Game's Popularity Analysis.
DOI: 10.5220/0012853800004547

In *Proceedings of the 1st International Conference on Data Science and Engineering (ICDSE 2024)*, pages 582-588
ISBN: 978-989-758-690-3

processed data from the brief records of more than 300,000 patients over a four-year period through a computerized CNN model to analyze the disease problems that may have occurred after them, which demonstrated that there is a great potential for machine learning in the medical field (Cheng, et al. 2016). Similarly, Chio et al. in the same year, utilized another deep learning approach RNN to predict the symptoms of heart failure in patients (Choi, et al. 2017). This again shows that machine learning has more application scenarios in the medical field. Moreover, apart from the medical field aspect, machine learning can also work on the prediction of energy utilization. Jun Wang et al. accurately predicted the output power of a photovoltaic power generation system using a combination of algorithms and neural networks, which provided a great deal of help to the researchers (Yu & Xu 2014). In addition to this, Azad, Md Shawmoon et al. combined the Theory of Planned Behavior (TPB) and machine learning methods to construct a new prediction model that can make predictions about consumer purchasing behavior on social platforms, through which they went to summarize what are the main factors that consumers care about in their minds. These principles and methods combined with the modeling approach have a greater auxiliary role in the management and sales of products.

The main objective of this paper's work is to find out the weights of these factors affecting the positive reviews of the game by processing the dataset from the Steam platform which contains multiple factors, and to analyze the data by finding a better model to ultimately give the game companies and others a way to understand the player's intention and make better decisions (Shawmoon, et al. 2023). In this paper, the authors use a variety of machine learning algorithms, including logistic regression, SVM, decision trees, random forests, XGBoost, and LightGBM, to process the sample dataset with these machine algorithmic models in turn, and to compare the differences in the performance of the above seven models by using the metrics of accuracy and area under the curve (AUC) in order to determine the most effective algorithm model. After the experiments it can be seen that the LightGBM model is the most effective among these models. Specifically, from the data, LightGBM has an accuracy of 85.62% and an AUC of 76.95%. This experimental result shows that LightGBM outperforms other models in these evaluation criteria. The model is intended to serve as a strategic tool for game developers and companies, and helps them to make quick and informed decisions, ultimately reducing investment risks. In addition, it provides valuable insights for community administrators to manage the player community more effectively, thus contributing to the overall maintenance and development of the gaming community.

2 METHODOLOGY

2.1 Dataset Description and Preprocessing

This study uses a dataset called "Steam Store Games" available on the Kaggle platform (Kaggle 2019). Collecting nearly 27,000 games from the Steam Store and SteamSpy APls, the dataset provides information on various aspects of the games in the Steam store, such as the type of game, number of owners, etc., developer and publisher information, game tags, prices, and other characteristics. After that, the authors fix a few problems with the dataset by removing rows that had ratings missing, modifying the types of columns, dealing with duplication, etc. Some games, for instance, require having their missing user rating data fixed before they could be analyzed.

2.2 Proposed Approach

This study uses a comparative systematic approach to determine the most effective model for predicting favourable reviews of game genres. A key aspect of paper data processing approach is the introduction of a crucial classification rule, where games receiving a positive rating above 90% are designated as "positive games." This categorization is instrumental in streamlining the dataset for a more focused analysis, particularly targeting games with favorable user reception. Firstly, the dataset is subjected to preprocessing and preliminary analysis, which is vital in uncovering significant correlations among the data. This process was important in determining what key data elements were needed for further in-depth analysis. Next, the authors used seven different machine learning predictive models for their analysis. Each model was rigorously evaluated using two performance metrics: accuracy and the AUC. To facilitate a thorough understanding of the effectiveness of each model, author visualizes the predictive results. Thirdly, the research involves a comparative analysis of these models. The aim is to pinpoint the model that most accurately predicts positive reviews within the specific context of game genres. This comparative approach is pivotal in ensuring a data-driven and systematic selection of the optimal predictive model, aligning with stated research objectives. The process is shown in the Figure 1.

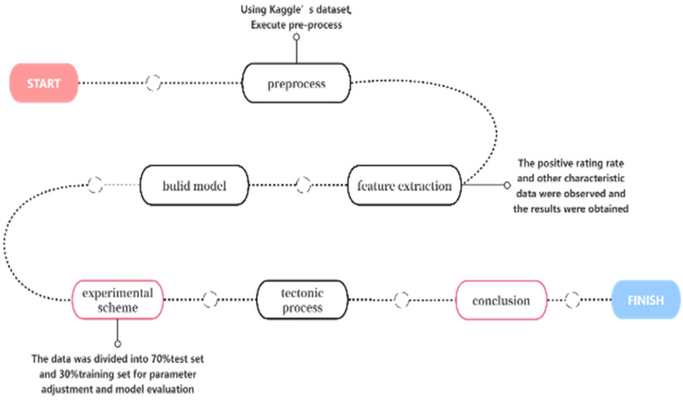
Using Kaggle' s dataset,
Execute pre-process
START
preprocess
bulid model
feature extraction
The positive rating rate and other characteristic data were observed and the results were obtained
experimental scheme
tectonic process
conclusion
FINISH
The data was divided into 70%test set and 30%training set for parameter adjustment and model evaluation

Figure 1: The process of the study (Picture credit: Original).

2.2.1 Traditional Single Model

Logistic Regression, commonly utilize in the context of binary classification tasks, functions to estimate the probability of an event's occurrence, ensuring the resultant output values lie within the 0 to 1 range. The core of logistic regression is the logistic function, usually embodied as a sigmoid function. The key to this functionality lies in transforming any real-valued input into a bounded output, especially one between 0 and 1, thus making the output a probability measure. The ability to map real numbers to probability values allows logistic regression models to make reasonable predictions about the likelihood of specific categories or events. Logistic regression can be used to create a clinical prediction model with binary outcomes (Maren, et al. 2019).

SVMs are recognized as better classification models and are particularly effective for working with high-dimensional data. The main function of an SVM is to determine an optimal boundary (called a hyperplane) that effectively separates data into different classes. The principle of SVM is to operate by maximizing the boundaries between different categories of data points. This method can effectively improve the accuracy of classification. One of the main features of SVM is the use of the "kernel trick", which improves mapping data into a high-dimensional space to solve linearly inseparable problems in the original space. This technique becomes more effective when the data is not linearly separable in the original feature space, allowing for more complex and flexible decision boundaries. SVM's adaptability to various data structures and its ability to process high-dimensional data make it an important tool in the field of machine learning and is widely used in this field. SVM models are also powerful in the medical field and can be used to predict the likelihood of common diseases such as diabetes and prediabetes (Yu, et al. 2010).

The decision tree model uses a tree structure in which each internal node represents a test based on a specific feature. This test can effectively distinguish data points based on different values of the specific feature. The branches of these nodes represent the various results of these tests, which ultimately form leaf nodes corresponding to categories in a classification tree or continuous values in a regression tree. The construction of a decision tree involves a recursive process of feature selection and segmentation. The purpose of this process is to create an efficient and accurate decision tree, which is suitable for predicting academic performance (Decision Tree-Based Predictive Models for Academic Achievement Using College Students' Support Networks, 2021). The purpose of this process is to divide the entire data set into Subsets, this criterion is often evaluated using metrics such as information gain or Gini impurity. Because the hierarchy of decision trees and their intuitive structure are easy to understand and explain, they can become a widely used tool in the field of machine learning.

2.2.2 Deep Learning Models

The main application point of CNN is to process image data, and it is widely used in fields such as image recognition, facial recognition and target detection. The underlying mechanism of CNN is to automatically and efficiently extract features from images through convolutional layers. These layers utilize convolution kernels or filters that are moved over the image to capture local features in different regions. Subsequently, further processing is performed using pooling layers (e.g., max pooling), with the aim of reducing the dimensionality of the data while highlighting the most salient features. The

CNN architecture also integrates fully connected layers, allowing the model to perform related tasks such as classification based on pre-extracted features. Using this specific architecture, CNN can perform tasks involving complex image data very efficiently. For example, CNN models can be used to predict the survival rate of pneumonia patients (Fahime, et al. 2021).

DNN is a more complex and powerful neural network architecture composed of multi-layer perceptrons. Therefore, DNN is very suitable for processing various data types and tasks, especially problems that are difficult to solve with simple linear models, such as speech recognition and natural language processing. At the heart of a DNN are multiple hidden layers, each of which contains a large number of neurons responsible for learning high-level data features. In a DNN, each neuron connection is assigned a weight and is subject to fine-tuning operations during the training phase. This tuning is achieved through a back-propagation algorithm, primarily to minimize prediction errors. DNNs require large amounts of data for effective training and are prone to overfitting problems. Still, the inherent complexity makes DNNs an important component when dealing with complex data analysis and pattern recognition tasks. For example, DNNs are used to predict diseases from laboratory test results (Jin, et al. 2021).

2.2.3 Ensemble Learning Model

Ensemble learning methods are mainly used here, namely Random Forest (RF), Extreme XGBoost, and LightGBM.

Multiple decision trees are used in RF, an ensemble learning technique, to lower variance and boost generalization performance. The training data and attributes for each tree are randomly sampled using bagging in RF, and the tree predictions are aggregated by majority vote or average. RF can handle both numerical and categorical variables and is resistant to noise and outliers. Nevertheless, training and prediction times for RF may be longer than for other techniques, and it might not be able to capture intricate nonlinear correlations and interactions between data.

Another ensemble learning technique, called XGBoost, builds an ensemble of decision trees in a stepwise manner by employing gradient boosting. By adding trees one at a time and fitting the preceding tree's negative gradients, often known as residuals, XGBoost optimizes a differentiable loss function. Regression and classification issues can be handled using XGBoost, along with automatically handling feature scaling and missing values. Moreover, XGBoost offers several regularization strategies to reduce overfitting and enhance accuracy, including shrinkage, subsampling, and pruning. However, hyperparameters might need to be carefully adjusted because XGBoost can be susceptible to noise and outliers.

LGBM is a framework for implementing the GBDT algorithm. It supports efficient parallel training with the advantages of faster training speed, lower memory consumption, and fast distributed processing of large amounts of data. LGBM has a wide range of application scenarios and can usually be used for tasks such as binary classification, multi-classification, and sorting. It completes 3 optimizations based on XGBoost.

$$LGBM = XGBoost + Histogram + GOSS + EFB \tag{1}$$

The core idea of the algorithm is a histogram-based decision tree learning algorithm, which discretizes continuous real feature values into different discrete values, and then uses the histogram algorithm to train and predict the discretized features.

The gradient boosting algorithm used is LGBM which minimizes the loss function by iteratively adding decision trees. Each new tree is constructed based on gradient information from the prediction residuals of all previous trees. This can be expressed as:

$$\hat{y}_i^{(t)} = \hat{y}_i^{(t-1)} + \rho \cdot h_t(x_i) \tag{2}$$

In formula (2), $\hat{y}_i^{(t)}$ is the predicted value after the t-th round of iteration, $h_t(x_i)$ is the x_i-prediction of the sample by the t-th tree, and ρ is the learning rate.

Gradient-based one-sided sampling (GOSS) is a method used to improve computational efficiency. It mainly focuses on those samples with the largest loss when calculating the gradient of samples. This method is represented by the following formula:

$$\{(x_i, y_i) \mid abs(g_i) > threshold\} \cup Random\ Sample\{(x_i, y_i) \mid abs(g_i) \le threshold\} \tag{3}$$

In formula (3), g_i is the gradient value of sample i, and threshold is the threshold used to select samples. Exclusive Feature Bundling (EFB) is one of the methods to reduce the number of features in high dimensional data. This strategy can be expressed by the following formula:

$$EFB = \oplus\{f_j \mid f_j \text{ is exclusive to } f_k, \forall k \neq j\} \tag{4}$$

In formula (4), ⊕ represents the bundling operation of features, and f_i represents a single feature. With fewer time and resources than XGBoost, LGBM can produce results that are on par with or even better than XGBoost.

3 RESULT AND DISCUSSION

In this study, author undertake a comprehensive exploratory data analysis of the Steam game dataset to elucidate the interrelations among various variables and to understand the dynamics governing the gaming market. Initially, author approach involves the construction of a heatmap. The heat map and heat bars are shown in Figures 2 and 3. This visual representation is instrumental in delineating the correlations existing among the dataset's variables, offering a clear and insightful perspective into the data's underlying structure and relationships.

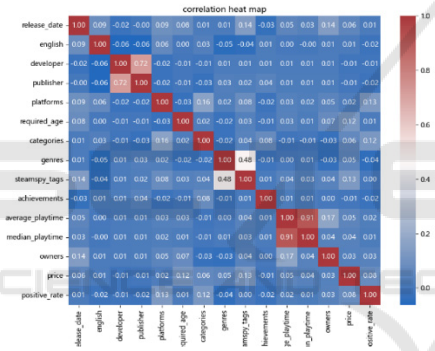

Figure 2: The correlation heat map of the data (Picture credit: Original).

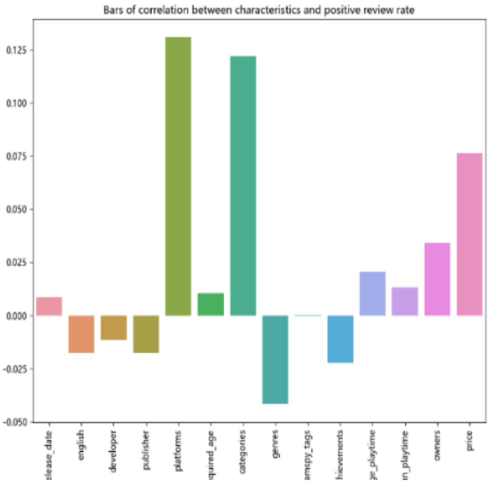

Figure 3: The bars of correlation between characteristics and positive review rate (Picture credit: Original).

The study goes on to explore in detail the impact of "gaming platforms" on the popularity of games. This analysis uses bar charts to determine the relationship between the number of platforms a game supports and player preference. A visual representation of the charts outlines the distribution of high (greater than 0.9) and low (less than or equal to 0.9) rated games across different platforms such as Windows, Mac and Linux (Figure 4 and Figure 5). It can be seen that the more platforms a game spends, the more favorable ratings become correspondingly. In addition, the study also explores the impact of both single-player and multiplayer game formats (Figure 6) as well as game categories such as action, adventure, and role-playing on players' choices (Figure 7). It can be seen that single-player games will have better favorable ratings; in terms of game genres, indie, casual, and action games also have higher favorable ratings. This multifaceted research approach provides a nuanced understanding of the factors that influence the appeal and success of games in a competitive gaming market.

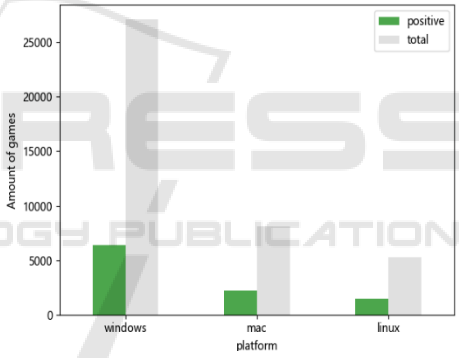

Figure 4: The percentage of games with positive ratings greater than 0.9 on different platforms (Picture credit: Original).

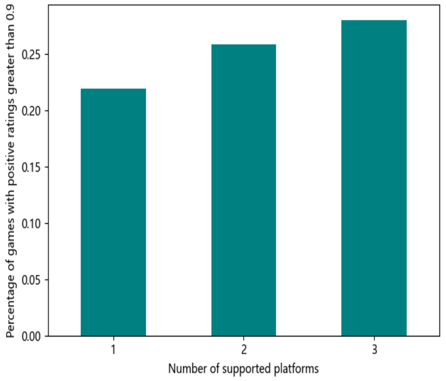

Figure 5: The number of supported platforms and percentage of games with positive ratings greater than 0.9 (Picture credit: Original).

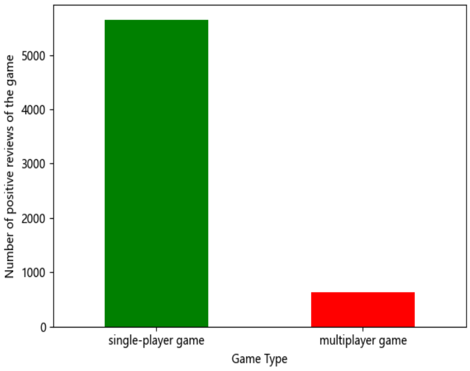

Figure 6: The percentage of positive reviews for single-player and multiplayer games (Picture credit: Original).

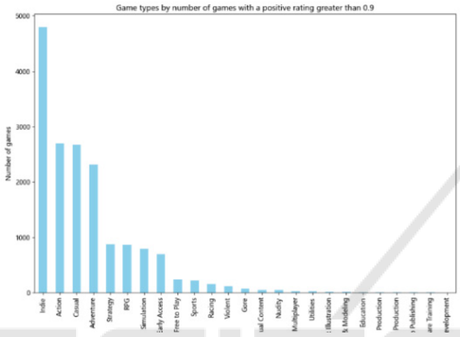

Figure 7: Game types by number of games with a positive rating greater than 0.9 (Picture credit: Original).

In this study, the authors also delve into the distribution of game prices (Figure 8). This allows for the division of game prices into different categories and examines the prevalence of games within each price band. It can be seen that as the price of the game increases, the critical acclaim of the game decreases. By categorizing and analyzing prices, it allows gaming companies to have a better understanding of pricing and aids them in making better decisions on pricing strategies.

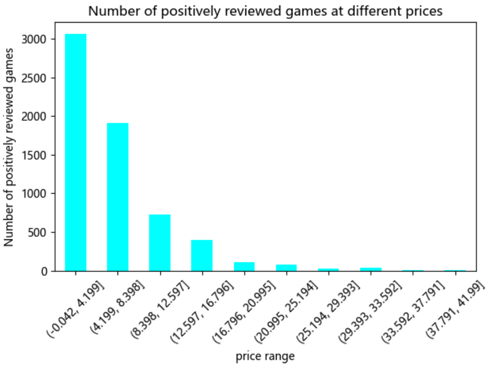

Figure 8: The number of positively reviewed games at different prices (Picture credit: Original).

After processing the data initially in this study, the relationship between the data features and the game's popularity can be seen more clearly. This facilitates the authors of this paper to further process the dataset using various algorithmic models, and we filtered out the data features that are more relevant to the game's popularity in order to improve the basic performance of various algorithms, so that the initial noise and error of the model can be effectively eliminated.

The Accuracy and AUC of the different models of this study are shown in Figure 9.

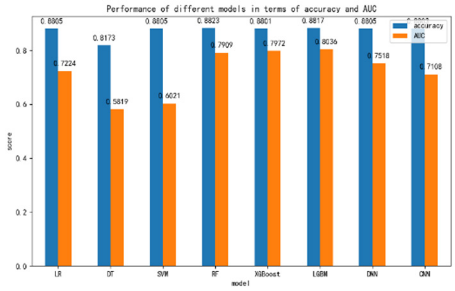

Figure 9: The performance of different models in terms of accuracy and AUC (Picture credit: Original).

After analyzing and comparing the algorithmic models among them, it is found that none of the traditional algorithmic models seem to perform very well, and it is found that there is a very significant difference between them by looking at the AUC metrics. In these data results, it can be seen that the LGBM model has an accuracy of 88.17% and an AUC of 80.36%, which is the best result compared to the other models. In addition, DNNs and CNNs also show a relative advantage in terms of AUC performance.

4 CONCLUSION

Overall, in order to find the key factors that affect the popularity of games on the steam platform, this study uses a variety of algorithms: logistic regression, SVM, decision tree, XGBoost, LGBM, DNN and CNN to analyze and construct the Steam store game data set. Modeling and optimization. Among them, LGBM can effectively handle large and complex data sets due to its high efficiency and high accuracy. The process of modeling involves tree-based algorithms for gradient boosting and complex relationships between various factors. The results show that platform, game type, tag, and price have a significant impact on popularity. Using this model, game developers have a clear understanding of the main factors of the popularity of games on the Steam platform. It would be useful for future research to apply the established analytical framework to a broader range of digital markets. This comparative analysis can provide valuable insights into the evolution of the gaming industry and changes in player preferences. Additionally, future research could focus on incorporating real-time data analytics that can ultimately increase a game's popularity by capturing the dynamics inherent in consumer behavior patterns and developing strategies to enhance popularity factors.

AUTHORS CONTRIBUTION

All the authors contributed equally and their names were listed in alphabetical order.

REFERENCES

arXiv preprint - Decision Tree-Based Predictive Models for Academic Achievement Using College Students' Support Networks, 2021, available at arXiv:2108.13947.

arXiv preprint - Predicting the Popularity of Games on Steam, 2021, available at arXiv:2110.02896.

D. Jin, et al. Scientific Reports, p. 7567 (2021).

E. Choi, et al. Journal of the American Medical Informatics Association, pp. 361-370 (2017).

E. Maren, et al, Journal of Thoracic Disease, 11, p. S574 (2019).

F. Yu and X. Z. Xu, Applied Energy, pp. 102-113 (2014).

J. M. Palma-Ruiz, A. T. Toukoumidis, S. E. González-Moreno, et al, Heliyon (2022).

K. Fahime, et al. Scientific Reports, p. 15343, (2021).

Kaggle - steam-store-games, 2019, available at https://www.kaggle.com/datasets/nikdavis/steam-store-games.

M. Shawmoon, et al. Plos one, p. e0296336 (2023).

W. Yu, et al. BMC medical informatics and decision making, pp. 1-7 (2010).

Y. Cheng, et al. "Risk prediction with electronic health records: A deep learning approach," in *Proceedings of the 2016 SIAM international conference on data mining*. Society for Industrial and Applied Mathematics, (2016), pp. 432-440.