IRSTI 55.01.05

Nurdaulet Anefiyayev[1], Dayana Anefiyayeva[2]
[1,2]Suleyman Demirel University, Kaskelen, Kazakhstan

PREDICTION OF THE BUSINESS ORIENTATION BASED ON ANALYSIS OF USER DESIRES

Abstract. Business is a well-established system aimed at making money and an activity that not everyone can succeed in. Business should work like clockwork that always works well with the most important - data. Only on the basis of data analysis can we make informed and rational decisions about where the business should develop, its methods became common in business when technologies for collecting and processing data developed. One of the main decisions that can be taken on the basis of data processing is to predict in which area it is better to develop a business. This paper aims to build a correct understanding of what customers are interested in based on feedback from customers. Improved machine learning solves this problem and algorithms like decision trees, neural networks, and regression are used to analyze issues and choose the best option - a foundation for the future strategy of the company.

Keywords: Customer opinion, feedback, customer behaviour prediction, income prediction, machine learning approaches.

***

Аңдатпа. Бизнес - бұл ақша табуға бағытталған жақсы қалыптасқан жүйе және кез келген адам табысқа жете бермейтін сала. Бизнес әрқашан сағат сияқты жұмыс істеуі керек және ең бастысы - деректермен. Деректерді талдау негізінде ғана біз бизнестің қай жерде дамуы керектігі туралы негізделген және ұтымды шешім қабылдай аламыз, оның әдістері бизнесте деректерді жинау және өңдеу технологиялары дамыған кезде кең таралған. Мәліметтерді өңдеу негізінде қабылданатын негізгі шешімдердің бірі – бизнесті қай салада дамыту тиімді екенін болжау. Бұл құжат тұтынушылардың пікірлері негізінде тұтынушылардың нені қызықтыратынын дұрыс түсінуге бағытталған. Жақсартылған машиналық оқыту бұл мәселені шешеді және шешімдер ағаштары, нейрондық желілер және регрессия сияқты алгоритмдер мәселелерді талдау және компанияның болашақ стратегиясының негізі ретінде ең жақсы нұсқаны таңдау үшін қолданылады.

Түйін сөздер: Тұтынушы пікірі, кері байланыс, тұтынушының мінез-құлқын болжау, кірісті болжау, машиналық оқыту тәсілдері

***

Аннотация. Бизнес — это отлаженная система, направленная на зарабатывание денег и деятельность, в которой не каждый может преуспеть. Бизнес должен работать как часы, которые всегда хорошо работают с самым важным — данными. Только на основе анализа данных мы можем принимать взвешенные и рациональные решения о том, куда должен развиваться бизнес, его методы стали распространены в бизнесе, когда развились технологии сбора и обработки данных. Одно из основных решений, которое можно принять на основе обработки данных, — спрогнозировать, в какой сфере лучше развивать бизнес. Эта статья направлена на построение правильного понимания того, что интересует клиентов, на основе отзывов клиентов. Усовершенствованное машинное обучение решает эту проблему, а такие алгоритмы, как деревья решений, нейронные сети и регрессия, используются для анализа проблем и выбора наилучшего варианта — основы будущей стратегии компании.

Ключевые слова: Мнение клиентов, обратная связь, прогноз поведения клиентов, прогноз доходов, подходы машинного обучения

I. Introduction

People always need advice and recommendations when faced with a choice: it could be the choice of restaurant, book or film, etc [1]. Before the spread of the global Internet, they asked their friends for advice in traditional ways. With the advent of web platforms such as blogs, forums, and social networks, the process of obtaining recommendations has become much easier. Companies have appeared that offer their services and have formed a customer flow. One of the most important tasks for companies is the task of predicting customer churn. Churn Rate is a business term that describes the rate at which customers leave a company. This is a key metric for the company, as acquiring new customers is much more expensive than retaining existing ones. Therefore, understanding how to keep customers engaged is the logical foundation for developing customer retention strategies.[1][2] As a result, businesses are looking for better technology to identify possible customer exits. In order to retain users and increase their number, the company must know how to grow and what to change; for this, it is necessary to analyze the actions of customers and take into account their wishes, otherwise, the business costs will increase, and marketing also will not help.

A company that provides payment services and other convenient simplified functions for people. Each year, the company sets strategic goals and chooses a path to help users and increase the number of loyal customers. In order for a company to grow and receive large profits, it is necessary to make the right assumptions and find the right ways to analyze user actions, and machine

learning prediction algorithms will help with this. Already presented algorithms and tools in the marketplace will not show best indicators, because each company is unique, and relying on its own data collected over years needs to be used upon the aim of the company. Having chosen the goal and path of the company according to the data, the company can also run startups, create and develop a new direction, a new project. This will help to focus in one direction on a single product because people want it. Usually, startups take risks when entering the market and in a competitive environment, but in our case, the company will support another branch.

The purpose of the research in this article is to create an application tool that allows you to visualize customer behaviour and predict the best business orientation for the company. This development will allow the company to react in time and maintain the interest of the client, as a result of which there will be a strengthening of relations and resistance to migration. The main objectives of the experiment are preparing a dataset for training and testing; research and transformation of data; choosing a machine learning algorithm with the best results; design and developing application tools.

This article is organized as follow: the next section Literature Review shows works of other researchers, followed by own research with data, and the last section shows results and conclusion, as well as limitations and future work.

II. Literature review

Several studies have been investigated to find the best customer approach and implications for business. Companies provide strategic meetings deciding to which category of business they need to pay more attention and focus. Since businesses are different, but the goal of all companies is to earn more money and loyal customers, need to observe the experiences of other workers, taking into account their failures and successes. Previous researches by other authors show us that the problem can be approached in different ways, which will be further described. The data represented in Table 1 demonstrates brief and important points of similar articles. One of the approaches sets a goal of the experiment to create a software tool that allows visualize customer behavior and predict his likely exit from the company. The authors used two data sets of clients: training for the period from January 2015 to June 2016 (18 months) and testing for the period from July 2016 to December 2016 (6 months) to conduct the experiment. Total number of unique customers whose data was used in the study: 500 with 16 attributes. The samples do not include data on new customers for the period under consideration. Based on the topic, the author identified several promising machine learning methods: neural networks, logistic regression, random forest, gradient boosting over decision trees. The best model was chosen based on the prediction results on the test sample. The final comparisons of the classifiers

were made using the Area under the ROC Curve (AUC) metric. The author's article showed the best results for the random forest model. [3].

Researchers of Creative Design LLP, which produces clothes for children, have found another solution. The enterprise was very young, and thus did not perform a full-fledged analysis of competitors; so there was a possibility that the offered goods would be of poor quality and not competitive. Hence, there is a need to create a tool within the limited budget of a small business that will help entrepreneurs in making the right decisions regarding the creation of certain goods or services. In short, they needed a system that would analyze the views of members of the Internet community on various topics of discussion. The author used two metrics (recall and precision) for the correctness of text classification. Text classification was divided into 4 classes: true positive, true negative, false positive, and false-negative. In order to classify authors used classification models such as n-gram, Naive-Bayes and kNN and for testing results were used cross-validation check. [4].

The findings of the another paper, is to process consumer feedback, opinions provided on the support cases to learn about different attributes that contributed to a positive or poor customer experience. Artificial neural networks are used to train on historic data available from customers and then provide classification guidelines on the current ongoing case compared to that. Artificial neural networks and Naive Bayes have different ways of treating and processing data; the author compared both of them with finite results and calculations. The data was taken from the TripAdvisor website, an old portal for hotel bookings, and has a listing of more than 200 hotels and around 30 destinations. The sentiment analysis done shows the accuracy of Naive Bayes is 72.06 % and Neural Network shows the accuracy as 80 % which is more than 7 % progress over the formal proposed model. [5].

Overall, the literature review reveals that machine learning algorithms have the potential to increase various aspects of business operations. Authors support the idea that using machine learning algorithms and tools demonstrated their ability to provide valuable insights and predictive capabilities. However, the successful adoption of machine learning algorithms in business requires careful consideration of several factors, such as data quality, model accuracy. By understanding these factors and using the latest advances in machine learning research, this paper can unlock new opportunities for growth and competitiveness in the market.

Table 1. Related works.

No	Research	Goals and objectives	Strategy / Approach	Performance	Limitation and Future Work
1	Meldebay M.A. and Sarbasova A.K.[4]	Need a system to analyze views of members of the Internet community on various topics of discussion.	Naive-Bayes, KNN, n-gramm	Cross validation algorithm to check classifiers.	show results of all classifier's
2	Li Huang , Gan Zhou, Jie Li, Shihai Yang[6]	realize excellent mechanisms of information interaction and economic, needed between power grid and users	BPNN, Markov, Similar day algorithm	error rate of Markov model is only-7%~8% compared to around 20% of similar day algorithm	prediction will be improved from two sides, including more detailed states of the appliance and the multiple order Markov model
3	Reznichenko E.V. and Matveev M.G.[3]	create a software tool to visualize customer behaviour and predict his exit	Neural network, Logistic regression, Random Forest, Gradient boosting	Precision, Recall, F1, AUC	Random forest showed best results, but for further works were chosen linear regression
4	Amit Ganesh Upadhye and A.C Lomte[5]	Process consumer feedback to learn about different attributes that contributed to positive or poor customer experience	Neural network, Naive Bayes	Compared to sentiments of customers, neural networks shows better results	Apply another data mining process on this kind of data which may or may not provide better results
5	Haihua Xie, Jingwei Yang, Carl K. Chang, Lin Liu[7]	Propose CRF methodology to provide quantitative exploration of a system-user interactions	CRF methods: divergent behaviour, goal transition, erroneous behaviour)	accuracy of goal inference is high (> 90%)	only able to capture users' emerging functional requirements, not non-functional

III. Methodology

Since business orientation prediction requires many aspects, the required data includes multiple datasets. The first dataset is a collection of reviews for describing a partner's service, which also includes rating and created date. The second dataset is events collected from the application for the last 5 years stored

in the Firebase system. Dataset exposes events to provide insight into what is happening in a company's application, such as user actions, system events, or errors. Events are text-based data with additional parameter values, used as filters in audience definitions that can be applied to every report. Dataset has 643 rows and 4 columns, as shown in Fig. 1., displaying a list of events, how many times they were called and amount of users who interacted.

```
events_df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 643 entries, 0 to 642
Data columns (total 4 columns):
 #   Column              Non-Null Count  Dtype
---  ------              --------------  -----
 0   Event Name          643 non-null    object
 1   Count               643 non-null    int64
 2   Users               643 non-null    int64
 3   Mark as conversion  643 non-null    bool
dtypes: bool(1), int64(2), object(1)
memory usage: 15.8+ KB
```

Figure 1. Events dataset.

```
terminals_df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 7793 entries, 0 to 7792
Data columns (total 7 columns):
 #   Column          Non-Null Count  Dtype
---  ------          --------------  -----
 0   terminal_id     7793 non-null   int64
 1   location_id     7793 non-null   object
 2   terminal_name   7793 non-null   object
 3   location_name   7793 non-null   object
 4   created_at      25 non-null     datetime64[ns]
 5   updated_at      25 non-null     datetime64[ns]
 6   status          7793 non-null   object
dtypes: datetime64[ns](2), int64(1), object(4)
memory usage: 426.3+ KB
```

Figure 2. Terminals dataset.

The next dataset is currently active terminals of the company. This dataset contains data such as terminal id, name, location id, name, and created updated dates. Overall, there are 7793 partners, and the detailed terminal data as average price, rating, reviews, cashback, opening hours, and other services as shown in Fig. 2. Using this dataset, reviews will be analyzed related to other parameters.

Data preprocessing.

A. Data cleaning. The first dataset describes reviews and they are very noisy, so they need to be cleaned and get rid of unused parts of data to get the best result. Data preprocessing steps, in this case, are:

- Converting all letters to lowercase
- Removing all punctuations
- Remove Unicode emojis
- Removing any kind of URL, link
- All numbers are removed
- All usernames are replaced by the word 'user'
- Lemmatizing - we remove the endings from the words and return the basic form of the word, known as the lemma.

B. Feature vectorizing. To make the calculation easy - transform text strings to vector numbers using CountVectorizer, then reconstruct the TF-IDF Vectorizer view to make it more convenient to use. As we want to try to compare several algorithms, tokenization of our data is used to get a bag of words.

C. Train, validation and test set. The last step is making three sets: one for training - reviews dataset, validation - for keeping model to have small error loss and last for the testing well-structured model.

Many classification algorithms in machine learning can be realized to find the final algorithm to be used in the model, as seen in other research in the Literature Review section. To get the right and best results a few ML algorithms need to be implemented in this paper too. To begin with, reviews will be analyzed and classified with a confusion matrix: True positive, true negative, false positive, false negative. Also, as a metric to predict text-based classification will be used - the F1 score with precision and recall, using the formula shown in Pic. 3. Afterwards, events will be analyzed within each category and branch. Several classifiers have been chosen in order to find the orientation of business based on user desires, such as recurrent neural networks (RNN), Decision Tree, XGBoost Classifier and Naive Bayes. Two of the algorithms from the list were used in other papers, but in this experiment, changed parameters will be used, as well as new methods. In the end, the research predicts the business amount of revenue for each category using Linear Regression, applying training and testing vectors from the previous section and providing them with chosen algorithms to train. All the listed methods are represented in Table 2.

Table 2. Methods and metrics

<table>
<tr><th></th><th>Algorithms / Models</th><th>Metrics</th></tr>
<tr><td rowspan="3">Analyzing review</td><td rowspan="3">Confusion Matrix (TP, TN, FP, FN)</td><td>F1-score</td></tr>
<tr><td>Precision</td></tr>
<tr><td>Recall</td></tr>
<tr><td rowspan="4">Classifying business category</td><td>RNN</td><td rowspan="2">Accuracy</td></tr>
<tr><td>XGBoost</td></tr>
<tr><td>Naive Bayes</td><td rowspan="2">RMSE</td></tr>
<tr><td>Decision Tree</td></tr>
<tr><td>Revenue prediction</td><td>Linear Regression</td><td>Accuracy</td></tr>
</table>

$$F1Score = \frac{2*presicion*recall}{precision+recall} \quad (1)$$

$$Recall = \frac{TP}{TP+FN} \quad (2)$$

$$Precision = \frac{TP}{TP+FP} \quad (3)$$

IV. Results

The results shown in Tables 3 and 4 for review analysis, listing the outcomes for the confusion matrix, by applying actual values and predicted values. Customers between times make mistakes in leaving ratings and reviews and evaluate them incorrectly, therefore, in order to understand the real value, we apply a confusion matrix and this will let us understand the effectiveness of the category where the review belongs. In this experiment, 30000 review values were used and analyzed. The values of the matrix indicate 71.8 percent of F1-score value, which is the harmonic mean of precision and recall and is a better measure than accuracy.

Table 3. Confusion matrix for reviews

		Actual values	
		True	False
Predicted values	Positive	14705	4051
	Negative	3740	7504

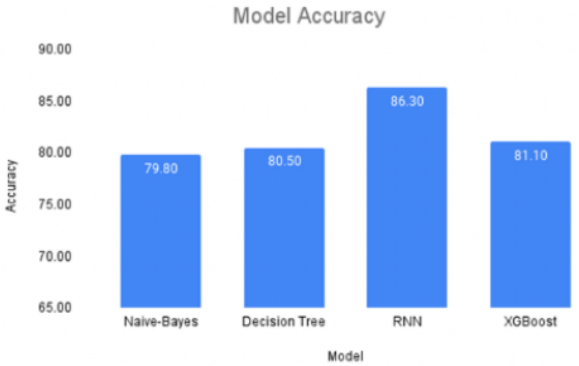
Table 4. Results for confusion matrix

Metric	Result
Precision	0.784
Recall	0.662
F1-score	0.718

In the following experiment, the effectiveness of the category is tested by their use in the application: the number of visits, transitions, time spent on pages of partner, count of locations and other parameters. There were partners who

refused to renew their contracts, they are also added to the list to find out if they got better or they were visited a few times only. In this process, the recurrent neural network showed a better result than others by more than 5%. The accuracy of the selected model shows 86.3% and with 0.09% mean error.

In the end, we calculated how profitable this or that category could be, a kind of funnel that shows the number of transactions for the total amount, the result of which category at the moment will generate income in the near future. The company can now rely more on the results of experiments and connect those partners with whom it will grow further and provide services.



Picture 4. Algorithm models accuracy

Picture 5. Algorithm models list

V. Conclusion and Future Work

Listing all of the above, we can affirm that the business can at least somehow relax and know where to move forward. The chosen solution explored

the potential application of the model (RNN) as an option that allows the company to operate on productive categories, in predicting and managing churn rate in businesses. Machine learning allows us to solve the most difficult problem in the process of evaluating algorithms - the problem of determining the appropriate algorithm. The developed model makes it possible to make decisions based on objective data. The analysis of texts in three languages slowed down the translation of data into numbers, so it was necessary to take into account all the morphological and grammatical foundations.

In addition, we were limited by a desktop computer that was not able to train models for a long time and faster. In future work, the forecasting model will be improved with the addition of monthly data, including more accurate conversion events, this will help to focus on developing guidelines for businesses to implement machine learning models that use monthly data in a practical and effective way, considering the unique challenges and opportunities presented by monthly data. Also replacing the local desktop with server-side architecture can potentially result in faster processing times, more efficient resource usage, and improved scalability.

References

1 J. Frankenfield, "Churn Rate", Investopedia, n.d., [Online]. Available: https://www.investopedia.com/terms/c/churnrate.asp [Accessed 5 January 2022]

2 S. Amaresan, "What Is Customer Churn", 2021, February 16, [Online]. Available: https://blog.hubspot.com/service/what-is-customer-churn [Accessed 5 January 2022]

3 E.V. Reznichenko, M.G. Matveev, "Development of a software for prediction of client chump using machine learning methods", Voronezh State University, 2017: pp. 287-292

4 M.A. Meldebay,A.K. Sarbasova, "Analysis of customer opinions based on machine learning", Al-Farabi Kazakh National University, 2017: pp. 360-364

5 A.G. Upadhye, A.C Lomte, "Customer opinion assessment using artificial neural networks and machine learning", JSPM's BSIOTR, Pune, 2020: pp. 1829-1835

6 L. Huang , G. Zhou, J. Li, S. Yang, "Short-term load prediction based on user behaviors analysis", The 4th International Conference on Smart Grid and Smart Cities, 2020: pp. 18-23

7 H. Xie , J. Yang , C. K. Chang , L. Liu, "A Statistical Analysis Approach to Predict User's Changing Requirements for Software Service Evolution", 2017: pp. 147-164