International Scientific Journal of Engineering and Management
Volume: 03 Issue: 06 | June – 2024
An International Scholarly || Multidisciplinary || Open Access || Indexing in all major Database & Metadata
ISSN: 2583-6129
DOI: 10.55041/ISJEM01931

A STUDY ON PREDICTIVE ANALYSIS FOR CUSTOMER RETENTION USING KNN-ALGORITHM

Mr. S.Narendran[1] Ms. V. P. Swetha [2]

1. II MBA Student, Panimalar Engineering College
2. Assistant Professor, Department of Master of Business Administration, Panimalar Engineering College

ABSTRACT

This study examines customer retention in a competitive sales environment, focusing on satisfaction, trust, and perceived value. Utilizing the K-Nearest Neighbors (KNN) algorithm, the research aims to predict and enhance customer loyalty and streamline monitoring processes. The KNN algorithm classifies data points based on proximity rather than decision trees. Findings reveal a retention rate of 81.12% and a churn rate of 18.87%, with customer activity, subtotal, and quantity identified as key retention factors. Despite the challenges of retaining customers amidst available alternatives, this study contributes to developing effective retention strategies in the sales sector.

INTRODUCTION

Retaining customers has become a top priority in today's competitive business environment, where customer expectations constantly evolve. Predictive analytics, particularly using the K-Nearest Neighbors (KNN) algorithm, offers a proactive approach to anticipate customer churn by leveraging historical data and statistical modeling. KNN's simplicity and effectiveness in handling nonlinear relationships make it a robust choice for customer retention analysis. This study involves preprocessing data, selecting significant features, and constructing predictive models to identify at-risk customers. Despite challenges like model interpretability and handling imbalanced datasets, KNN provides valuable insights into customer behavior. By preemptively identifying churn risks, businesses can tailor retention strategies, enhancing loyalty and profitability.

ISSN: 2583-6129
DOI: 10.55041/ISJEM01931

NEED OF THE STUDY

- Anticipating customer churn by analysing historical data patterns.
- Identifying key factors influencing customer retention for targeted interventions.
- Enhancing marketing strategies through personalized recommendations.
- Optimizing resource allocation by prioritizing high-risk customers.
- Improving overall customer satisfaction and loyalty through proactive measures.

OBJECTIVES OF THE STUDY

SECONDARY OBJECTIVES:

- To study the predictive analytics in customer retention using KNN algorithm.
- To analyse customer data to identify distinct segments and preferences, enabling targeted retention strategies.
- To develop predictive models to forecast churn likelihood and proactively address attrition risks.
- To evaluate satisfaction levels through sentiment analysis to improve service and retention.
- To optimize loyalty programs with tailored rewards based on predictive analytics insights.
- To Implement personalized communication and interventions to mitigate churn risks and foster long-term relationships.

SCOPE OF THE STUDY

Predictive analytics with KNN in customer retention starts with gathering diverse data sources like transactions and behaviors. Feature selection pinpoints key factors for churn prediction. KNN develops a balanced, accurate, and interpretable model. Real-time predictions drive proactive retention strategies like incentives and personalized communications. Continuous monitoring optimizes model performance for dynamic business environments. Overall, it empowers businesses to proactively manage customer attrition risks.

REVIEW OF LITERATURE

Siegel, Eric (2019). Predictive Analytics: The Power to Predict Who Will Click, Buy, Lie,or Die (1st ed.). Wiley. ISBN 978-1-1183-5685-2.

ISSN: 2583-6129
DOI: 10.55041/ISJEM01931

Predictive analytics is a form of business analytics applying machine learning to generate a predictive model for certain business applications. As such, it encompasses a variety of statistical techniques from predictive modeling and machine learning that analyze current and historical facts to make predictions about future or otherwise unknown events. It represents a major subset of machine learning applications; in some contexts, it is synonymous with machine learning.In business, predictive models exploit patterns found in historical and transactional data to identify risks and opportunities. Models capture relationships among many factors to allow assessment of risk or potential associated with a particular set of conditions, guiding decision-making for candidate transactions.

Fitzpatrick, Dylan J.; Gorr, Wilpen L.; Neill, Daniel B. (2019-01-13). "Keeping Score: Predictive Analytics in Policing"

Predictive analytics in policing is a data-driven approach to (*a*) characterizing crime patterns across time and space and (*b*) leveraging this knowledge for the prevention of crime and disorder. This article outlines the current state of the field, providing a review of forecasting tools that have been successfully applied by police to the task of crime prediction. We then discuss options for structured design and evaluation of a predictive policing program so that the benefits of proactive intervention efforts are maximized given fixed resource constraints. We highlight examples of predictive policing programs that have been implemented and evaluated by police agencies in the field. Finally, we discuss ethical issues related to predictive analytics in policing and suggest approaches for minimizing potential harm to vulnerable communities while providing an equitable distribution of the benefits of crime prevention across populations within police jurisdiction.

Coker, Frank (2020). Pulse: Understanding the Vital Signs of Your Business (1st ed.). Bellevue, WA: Ambient Light Publishing. pp. 30, 39, 42, more. ISBN 978-0-9893086-0-1.

Predictive analytics is a set of business intelligence (BI) technologies that uncovers relationships and patterns within large volumes of data that can be used to predict behavior and events. Unlike other BI technologies, predictive

ISSN: 2583-6129
DOI: 10.55041/ISJEM01931

analytics is forward-looking, using past events to anticipate the future. Predictive analytics statistical techniques include data modeling, machine learning, AI, deep learning algorithms and data mining. Often the unknown event of interest is in the future, but predictive analytics can be applied to any type of unknown whether it be in the past, present or future. For example, identifying suspects after a crime has been committed, or credit card fraud as it occurs. The core of predictive analytics relies on capturing relationships between explanatory variables and the predicted variables from past occurrences, and exploiting them to predict the unknown outcome. It is important to note, however, that the accuracy and usability of results will depend greatly on the level of data analysis and the quality of assumptions.

Spalek, Seweryn (2019). Data Analytics in Project Management. Taylor & Francis Group, LLC. "Machine learning, explained". MIT Sloan. Retrieved 2022-05-06.

Machine learning is behind chatbots and predictive text, language translation apps, the shows Netflix suggests to you, and how your social media feeds are presented. It powers autonomous vehicles and machines that can diagnose medical conditions based on images. When companies today deploy artificial intelligence programs, they are most likely using machine learning so much so that the terms are often used interchangeably, and sometimes ambiguously. Machine learning is a subfield of artificial intelligence that gives computers the ability to learn without explicitly being programmed.“In just the last five or 10 years, machine learning has become a critical way, arguably the most important way, most parts of AI are done,” said MIT Sloan professorThomas W. Malone, the founding director of the MIT Center for Collective Intelligence. “So that's why some people use the terms AI and machine learning almost as synonymous … most of the current advances in AI have involved machine learning.”

Eckerson, Wayne, W (2021). "Predictive Analytics. Extending the Value of Your Data Warehousing Investment" .

Predictive analytics can identify the customers most likely to churn next month or to respond to next week’s direct mail piece. It can also anticipate when

ISSN: 2583-6129
DOI: 10.55041/ISJEM01931

factory floor machines are likely to break down or figure out which customers are likely to default on a bank loan. Today, marketing is the biggest user of predictive analytics with cross-selling, campaign management,customer acquisition, and budgeting and forecasting models top of the list, followed by attrition and loyalty applications

RESEARCH METHODOLOGY

The methodology section of a research paper delineates the systematic approach employed to investigate and address the research problem. It serves as a blueprint guiding the selection, processing, and analysis of information pertinent to the study. Research design, encompassing the overall strategy to integrate various study components coherently, is fundamental to ensuring the efficacy of addressing the research problem. In quantitative research, numerical data is systematically collected and analyzed through techniques like surveys, experiments, or observations to elucidate patterns or relationships. Sampling methods such as random sampling ensure the fair representation of the entire population, crucial for generalizing findings. Primary data, collected firsthand, and secondary data, sourced from existing resources, offer complementary insights. In this study, primary data from client feedback was collected using a rating system, while secondary data from journals and books provided additional support. Tools like Python, NumPy, pandas, matplotlib, seaborn, scikit-learn, statistical models, Jupyter Notebook, and Bokeh were instrumental in data analysis and visualization, facilitating comprehensive exploration and interpretation of findings.

ANALYTICS RESEARCH DESIGN

Analytical research involves the systematic collection and analysis of numerical data to understand phenomena, patterns, or relationships. This method employs structured data collection techniques like surveys, experiments, or observations to gather information objectively. By quantifying variables and using statistical and mathematical analyses, researchers aim to test hypotheses, generalize findings, and make predictions. Quantitative research provides precise measurements and statistical evidence, allowing for rigorous evaluation of relationships between variables. Its quantitative nature facilitates the identification of patterns and trends, making it a powerful tool for exploring complex phenomena in various fields, from social sciences to natural sciences and beyond.

International Scientific Journal of Engineering and Management
Volume: 03 Issue: 06 | June – 2024
An International Scholarly || Multidisciplinary || Open Access || Indexing in all major Database & Metadata
ISSN: 2583-6129
DOI: 10.55041/ISJEM01931

RANDOM SAMPLING

Random sampling ensures that every member of the population has an equal opportunity tobe included in the sample, thus providing a fair and representative representation of the entire population. This method involves selecting individuals entirely by chance, without any bias or preference towards specific characteristics. By randomly selecting samples from the population, researchers can minimize the risk of introducing systematic errors or biases into their study. This approach is crucial for generalizing findings from the sample to the entire population, as it ensures that each member of the population has an equal chance of being included, thereby increasing the validity and reliability of the research outcomes.

JUPYTER NOTEBOOK

Jupyter Notebook is a tool for conducting such analyses, especially when it comes to machine learning algorithms like K-Nearest Neighbors (KNN) for predictive analysis.

DATA ANALYSIS AND INTERPRETATION

CHURN PREDICTION:

OLS Regression Results			
Dep. Variable:	Churn	R-squared:	0.008
Model:	OLS	Adj. R-squared:	0.004
Method:	Least Squares	F-statistic:	0.001387
Date:	Tue, 16 Apr 2024	Prob (F-statistic):	0.970
Time:	15:17:29	Log-Likelihood:	-163.22
No. Observations:	276	AIC:	330.4
Df Residuals:	274	BIC:	337.7
Df Model:	1		
Covariance Type:	no robust		

	coef	std err	t	P>|t|	[0.025	0.975]

International Scientific Journal of Engineering and Management
Volume: 03 Issue: 06 | June – 2024
An International Scholarly || Multidisciplinary || Open Access || Indexing in all major Database & Metadata
ISSN: 2583-6129
DOI: 10.55041/ISJEM01931

const	0.2564	0.035	7.291	0.000	0.187	0.326
Price	2.365e-08	6.35e-07	0.037	0.970	1.23e-06	1.27e-06

Omnibus:	58.113	Durbin-Watson:	2.219
Prob(Omnibus):	0.000	Jarque-Bera (JB):	63.501
Skew:	1.111	Prob(JB):	1.63e-14
Kurtosis:	2.234	Cond. No.	7.37e+04

Inference: The regression results for churn prediction suggest that the model has limited expl anatory power, as indicated by the low R-squared value of 0.008. This implies that only a neg ligible proportion of the variance in churn can be explained by the independent variable, price. The coefficient for the price variable is not statistically significant, with a p-value of 0.970, s uggesting that price does not have a significant linear relationship with churn. The intercept(constant) term has a coefficient of 0.2564, indicating the predicted churn rate when the price is zero, though this is not practically meaningful in this context. Overall the model, based solely on price, is not effective in predicting churn in this dataset, and other factors beyond price likely influence churn behavior.

SATISFACTION LEVEL PREDICTION:

OLS Regression Results			
Dep. Variable:	Churn	R-squared:	0.018
Model:	OLS	Adj. R-squared:	0.014
Method:	Least Squares	F-statistic:	4.974
Date:	Tue, 16 Apr 2024	Prob (F-statistic):	0.970
Time:	15:17:29	Log-Likelihood:	-155.40
No. Observations:	276	AIC:	314.8
Df Residuals:	274	BIC:	322.0
Df Model:	1		
Covariance Type:	no robust		

ISSN: 2583-6129
DOI: 10.55041/ISJEM01931

	coef	std err	t	P>|t|	[0.025	0.975]
const	1.8076	0.034	52.882	0.000	1.740	1.875
Price	1.377e-08	6.17e-07	2.23	0.027	2.59e-06	1.61e-06

Omnibus:	58.113	Durbin-Watson:	2.219
Prob(Omnibus):	0.000	Jarque-Bera (JB):	63.501
Skew:	1.111	Prob(JB):	1.63e-14
Kurtosis:	2.234	Cond. No.	7.37e+04

Inference: The regression analysis shows a modest explanatory power for satisfaction level, with an R-squared value of 0.018, indicating that around 1.8% of the variance in satisfaction level is explained by price. The coefficient for price is statistically significant (p = 0.027), revealing a weak negative linear relationship with satisfaction level. Specifically, for each unit increase in price, satisfaction level decreases by approximately 1.377e-06 units. The intercept term is 1.8076. In summary, while price does influence satisfaction to some extent, other factors likely play a more substantial role in determining satisfaction levels in this dataset.

CLUSTER ANALYSIS:

Cluster	Month	Customer	Model	Projeccts	Age	Gender
0	6.386	5.416	3.00	20.535	34	1.891
1	6.593	5.692	2.92	13.385	30	1.725
2	6.619	5.964	2.94	13.607	31	1.630

Cluster	Price	Satisfaction Level	Tenure	Churn	Usage frequency	Payment Delay
0	35216.940	1.891	50.376	0.257	20.42	16.891
1	37011.329	1.725	26.538	0.472	11.09	22.252
2	37717.130	1.630	16.857	0.023	15.607	6.250

ISSN: 2583-6129
DOI: 10.55041/ISJEM01931

Cluster	Usage frequency	Payment Delay
0	20.42	16.891
1	11.09	22.252
2	15.607	6.250

Inference: Cluster 0 represents older users with a higher average age of 34.37 years, longer tenure of 50.38 units, and a correspondingly higher price point averaging at 35,216.94 units. They exhibit moderate usage frequency (averaging 20.43 times) and a lower churn rate of 25.74%. In contrast, Cluster 1 comprises younger users with an average age of 29.93 years, shorter tenure (25.54 units), and a slightly lower price point averaging at 37,011.33 units. Their usage frequency is lower (averaging 11.10 times) with a higher churn rate of 47.25%. Cluster 2 demonstrates characteristics falling between Clusters 0 and 1, with an average age of approximately 30.81 years, moderate tenure (16.86 units), and a price point similar to that of Cluster 1. Notably, they exhibit higher satisfaction and a significantly lower churn rate of 2.38%. Understanding these distinctions enables the formulation of targeted strategies tailored to each cluster's unique needs, potentially enhancing customer retention and satisfaction levels.

PREDICTIVE MODELLING

Cross-Validation ROC AUC Scores	0.87052342, 0.80588235, 0.79411765, 0.84411765, 0.85
Mean ROC AUC Score	0.8329282126073568

Inference: The cross-validation ROC AUC scores, ranging from 0.794 to 0.870, demonstrate consistent performance across data splits. With a mean ROC AUC of about 0.833, the model effectively distinguishes between classes, indicating reliability. While it's a strong indicator, other metrics and validation on unseen data are crucial for real-world applicability.

FINDINGS

- Customer demographics analysis indicates significant gender-based variations in

International Scientific Journal of Engineering and Management
Volume: 03 Issue: 06 | June – 2024
An International Scholarly || Multidisciplinary || Open Access || Indexing in all major Database & Metadata
ISSN: 2583-6129
DOI: 10.55041/ISJEM01931

average age, price distribution, satisfaction levels, tenure, and usage frequency. Females tend to have higher average ages, lower average prices, higher satisfaction levels, longer tenures, and slightly higher usage frequency compared

ISSN: 2583-6129
DOI: 10.55041/ISJEM01931

to males. Understanding these gender-specific differences is crucial for tailored marketing strategies and retention efforts.

- Churn analysis based on gender, age, satisfaction level, and tenure highlights distinct patterns in customer attrition rates. Females exhibit lower churn rates compared to males, with notable variations observed across different age groups, satisfaction levels, and tenure ranges. Addressing these variations can inform targeted interventions to improve customer retention and satisfaction, contributing to long-term business success.

SUGGESTIONS

- Customize experiences for male and female customers based on their unique preferences and behaviors to boost satisfaction and retention rates.
- Mitigate churn by targeting retention efforts towards specific demographic factors like age, satisfaction levels, and tenure. Utilize loyalty programs and improved customer support to foster long-term loyalty.

CONCLUSION

To optimize customer satisfaction and retention, businesses must adopt a comprehensive strategy. This involves understanding gender-specific preferences, analyzing churn patterns across demographics, refining pricing strategies, enhancing service delivery, and utilizing data-driven decision-making. Tailoring strategies to meet the unique needs of male and female customers, targeting retention efforts, and balancing revenue generation with customer satisfaction are crucial. Prioritizing service improvements and efficient payment processing, along with leveraging data insights, fosters long-term loyalty and growth. Continuous monitoring, proactive issue resolution, and adaptation to market dynamics are vital for sustained competitiveness and success.

ISSN: 2583-6129
DOI: 10.55041/ISJEM01931

ANNEXURE

BIBLIOGRAPHY:

BOOKS REFFERED :

- "Predictive Analytics: The Power to Predict Who Will Click, Buy, Lie, or Die" by Eric Siegel, Publisher: Wiley, Year: 2013
- "Practical Predictive Analytics: Using Python and Machine Learning" by Vishal Chaurasia, Publisher: Apress, Year: 2018
- "Marketing Analytics: Data-Driven Techniques with Microsoft Excel" by Wayne L. Winston, Publisher: Wiley, Year: 2016

JOURNALS REFFERED :

1. Finlay, Steven (2019). Predictive Analytics, Data Mining and Big Data. Myths, Misconceptions and Methods(1st ed.). Basingstoke: Palgrave Macmillan. p. 237. ISBN 978-1137379276.
2. Spalek, Seweryn (2019). Data Analytics in Project Management. Taylor & Francis Group, LLC.
3. "Machine learning, explained". MIT Sloan. Retrieved 2022-05-06.
4. Jump up to: Kinney, William R. (2020). "ARIMA and Regression in Analytical Review: An Empirical Test". The Accounting Review. 53 (1): 48–60. ISSN 0001-4826. JSTOR 245725.
5. "Introduction to ARIMA models". people.duke.edu. Retrieved 2022-05-06.
6. "6.4.3. What is Exponential Smoothing?". www.itl.nist.gov. Retrieved 2022-05-06.
7. "6.4.2.1. Single Moving Average". www.itl.nist.gov. Retrieved 2022-05-06.

WEBSITES REFFERED :

- https://scholar.google.com
- https://www.mendeley.com/
- https://www.researchgate.net/
- https://www.mcdermott.com/Home