Dean&Francis
ISSN 2959-6130

Does AI-based sentiment analysis predict consumer purchasing behavior in subscription services?

Fu Su

1. Abstract

This essay investigates the use of AI-written text sentiment analysis as a predictor of consumer churn in the digital streaming economy. The authors employed over 210,000 records of users of Netflix to generate the primary sentiment scores of the raw reviews left by the user by implementing a Natural Language Processing (NLP) algorithm. These were then inputted into a Logistic Regression model that predicted binary cancellation possibilities.

Despite the fact that preliminary values of the Weight of Evidence (WOE) factor have allowed establishing a mathematical relationship between the presence of extremely negative mood and subscription termination, the predictive factor has a final Area Under the Curve (AUC) prediction of 0.4944, which is analogous to a shot in the dark. The extremely low Total Information Value (0.0014) demonstrates that the human behaviour in purchasing is highly multidimensional. It is discovered that, as much as sentiment analysis is applied in the establishment of the overall brand health, text sentiment cannot be applied mathematically to engineer a commercially viable churn prediction system without integrating with other behavioral metrics.

Keywords: Artificial Intelligence, Sentiment Analysis, Consumer Churn, Subscription Economy, Logistic Regression, Natural Language Processing, Predictive Modelling.

2. Introduction

The world market in the last couple of years experienced the paradigm shift to the subscription economy of colossal proportions. The entertainment sector has been altered by digital streaming; e.g. Netflix; they have redesigned the one-time purchase model to the recurrent service-based revenue models. However, this is a radical transition that has introduced a business dilemma that is significant to the new business

ISSN 2959-6130

termed as customer churn which has experienced the rate at which customers end their active subscriptions. The customer retention in the case of subscription based, business is important in the long term profitability, as it is always expensive to acquire a new subscriber, relative to retaining an existing one. This, in turn, has made finding what the signals of consumer discontent are early enough and accurately forecasting churn before it occurs to become the primary objective of data scientists and marketing strategists operating in the digital media sector.

The modern finance and marketing sectors have turned to the application of Artificial Intelligence (AI) and Natural Language Processing (NLP) technologies in their search to prevent consumer churn increasingly. One of the most evident technological changes in the field, as of today, is AI-based sentiment analysis. It is accomplished by executing machine learning algorithms so as to automatically evaluate and quantify the emotional polarity, i.e. highly negative to highly positive, implicit in unformatted text-based data, such as customer feedback, comment cards and social media descriptions of information or products. The application of predictive mathematical models, such as the Logistic Regression, to compute the probability of binary outcomes, e.g. the probability of an individual credit risk or the probability of default under the condition of the historical data properties is not new in modern finance. Marketing industries have been fruitfully engaged in the implementation of these predictive algorithms and they gauge the well-being of brands and forecast consumer purchasing patterns. Based on mathematical categories of the emotional tone of the consumers, the companies attempt to perform selected data-based reactions such as a special discount or a personal customer care in order to save a failing consumer relationship.

Even though sentiment analysis application in e-commerce in general has become a very common practice, predictive value of this specific task defining the binary outcome of subscription renewal or cancellation has to be effectively statistically investigated. This essay closes this gap by applying the mathematical modeling method in consumer buying behavior in the streaming sector. Referring to a group of reviews left by the consumers and existing subscriptions of Netflix, this study will result in the primary sentiment value with the help of an NLP algorithm and feature analysis, binning, and a Logistic Regression model to test the predictive potentials in the text information. Based on the need to understand the relationship between text-based consumer emotion and the final purchase decision, this research will test the unambiguous hypothesis in the following way: The statistical significance of the lower AI-generated sentiment scores in user reviews as a predictor of the cancelled subscription status in streaming consumers. This question will be addressed in the project as it will test this hypothesis to understand whether AI sentiment analysis can be a perfect and adequate business solution to forecast consumer purchasing behavior in the subscription economy.

3. Literature Review

Introduction to the Review

The critical review of the scholarly literature should be presented, and the main data and mathematical models developed during the course of the current research should be offered. This research review will be carried out focusing on two academia themes. Firstly, it will examine the psychological and economic incentives of consumer churn and consumer buying in the fast-expanding digital subscription economy. Second, it will discuss how the use of data science (specifically, Natural Language Processing (NLP) and predictive statistical models, including the Logistic Regression), have long been applied to estimate human behavior and predict binary outcome trends in markets, such as finance and e-commerce. The combination of the two studies will assist this review in bringing about the theoretical background that will underpin the investigation and also gaps existing currently in predictive consumer modeling.

Existing Studies on Consumer Churn

The desire of digital entertainment has experienced a shift in structural change of a transactional buying model to Subscription Video on Demand (SVOD) model. Whereas this ensures the sustainability of the income of business models like Netflix, it poses the ever-present business threat of churn. Thus, the causes of the psychological and economic essence of these services cancellation by consumers have become the topic of a recent scholarly literature.

The new study claims that subscription fatigue and choice overload are said to be the initial factors that provoke the cancelation. Longo (2021) reports that the content array of the SVOD regularly grows exponentially, and the situation is also quite complex, which leads to the occurrence of choice overload when one is overwhelmed by the varieties of options instead of being satisfied with them. Equally, Nguyen (2025) has described subscription fatigue as mental and financial stress experienced by consumers when they are left to deal with a wide range of recurring payments that causes decision exhaustion and disorderly cancellations.

Furthermore, another significant aspect of the so-called perceived value is also illuminated with the assistance of the research. The general utility of a streaming service according to Srivastava et al. (2025) evaluates the price-quality relationship against brand image, which

directly dictates the willingness of a consumer to get subscribed to the service or not and the extent to which he or she retains it.

However, there is a very significant fiction in the instrumental application of this sociological and economic study. Although these studies are useful in determining the overall, high-level causes of demographics churn (fatigue, choice overload and dwindling value perception), there is not a lot of practical information provided to a business as to when a specific person is going to cancel his or her subscription. Theoretical explanations of churn are no longer enough to inform customer retention in contemporary context and the businesses must be informed by proactive and predictive mathematical models that are capable of forecasting consumer purchasing behavior in near future.

Predictive AI and Statistical Models

The organizations need to apply the predictive technology to focus on the psychological drive of subscription burning in an operational way. Recent information suggests the growth in the application of machine learning algorithms in advanced customer profiling in the streaming sector (Tanuwijaya et al., 2021). It subsequently logically follows that the present and future marketing can apply Natural Language Processing (NLP) algorithms to conduct automatic sentiment analysis, such as TextBlob. These algorithms are applied to large collections of raw consumer textual reviews of consumers as subjective human emotion being objectified into mathematical polarity scores.

When the emotion of consumers is quantified, it must be transformed into actionable intelligence in predictive statistical models. The industry standard in the binary classification problem such as in the case of whether a consumer will cancel (1) or renew (0) his subscription is Logistic Regression. Previous risk assessment studies by Martin (1977) and Ohlson (1980) had indicated that the Logistic Regression is especially effective in binary data when the probability is logistically distributed since cap-and-edge effects of probability (within a limit between 0 and 1) are included in the data. The same seminal models that have proven useful in extrapolating future bankruptcy and credit failures of corporations through the application of weighted financial measures are applied by the data scientists of the modern day to find out the churn predictabilities by use of consumer sentiment indicators. In addition, to critically assess the validity of these algorithms, the present mathematics literature employs the Receiver Operating Characteristic (ROC) curve and the Area Under the Curve (AUC) value to be the final measures of assessing the predictive ability and accuracy of a model in totality.

Identifying the Literature Gap

The intersectional location is still lacking, even though the literature already offers the capacity to explain the psychological reasons of subscription fatigue, as well as prove the usefulness of predictive statistical models. Currently AI sentiment analysis has been applied widely towards computing customer behavior and brand performance within the general e-commerce context, such as Amazon product review. Similarly, the official industry calculation of binary probabilities standard is the Logistic Regression with the use of financial services to assess and quantify credit default risk.

However, one can identify a huge gap in scholarly literature that explores the hypothesis of whether the text sentiment produced by AI is mathematically predictive of the binary outcome of subscription churn in the digital streaming economy. This essay bridges this specific gap in literature. This paper will involve a Logistic Regression model in a mathematical test of predictability of text sentiment using primary sentiment scores on a sample of Netflix user reviews. Lastly, this primary data research would determine whether natural language processing can be utilized well to reasonably predict whether a consumer will cancel or renew his subscription and whether it can be utilized well as a commercial instrument by itself.

4. Dataset Introduction and Methodology

In this study, a large and natural dataset was required to support the hypothesis of the possibility of using sentiment scores produced by artificial intelligence to predict subscription cancellations. The raw data was acquired using a full large scale open source dataset that Kaggle, Inc. is supplying Netflix user behavior which had over 210000 records of individuals. The raw material was especially cut out of two relational files, one of the user files, wherein the demographic information was retrieved along with the vital is-active subscription (is-active) flag and a review file wherein the raw textual feedback (review-text) left by the same consumers was retrieved. This information at large scale provided the statistical power necessary to form a powerful predictive model.

The adopted approach was a quantitative, data-driven one rather than a conventional one, i.e. a set of surveys by disinterested products consumers. Considering information related to the application of the questionnaire surveys within the context of the stage academic essay, recent analysis reports indicate that the use process is highly problematic, likely to subjective bias, and highly limited with regard to the usage of sample size. By substituting the surveys with a mathematical modeling system, which in the scenario of this project is the mirroring of the Logistic Regression algorithms typically used in financial credit risk assessment to predict binary outcomes, statistical validity, accuracy and objectivity of the consumer

ISSN 2959-6130

purchasing behaviour analysis will increase significantly in this project.

Whereas the initial Kaggle reviews dataset already constitutes calculated sentiment and sentiment_score variables, the adoption of the latter would have left this research to the realm of a pure secondary data analysis. These sentiment columns that exist were intentionally omitted in the information to increase the level of academic rigor and generate primary data.

Instead, the free Natural Language Processing (NLP) program, which is built on the Python library TextBlob, was adopted. They had to apply this NLP tool to the raw review text strings in the dataset to evaluate the emotional polarity of the text automatically. This calculation was done and was able to form a new, defined variable: the primary-sentiment-score which rated all the reviews on a continuous, numerical scale of -1.0 (very negative) to +1.0 (very positive).

This study will make sure that the further statistical analysis and predictive modeling of these newly emerging primary data can be grounded only in the data generated after purging of the secondary sentiment data, and development of these new measures of polarity on the basis of the raw text. This now fined sentiment score shall ultimately be determined by using the binary is active status of the users to determine the actual mathematical prediction of the consumer churn.

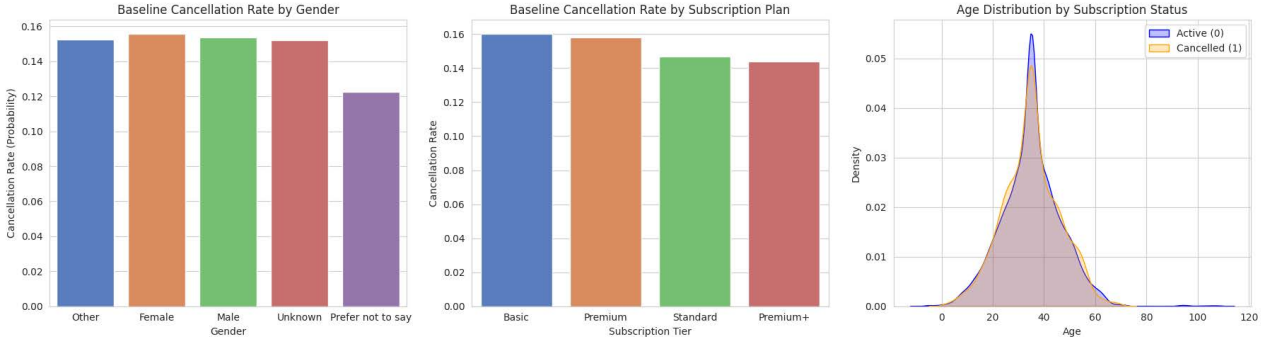
1. Data Preprocessing and Feature Analysis

The first step of the preprocessing of the data elements involved the initial loading of the data elements in a Python computational machine comprising of the Kaggle raw users and reviews files using the pandas data manipulation library. The common user id parameter was then used to combine the two datasets, and in effect, the individual demographic profile and subscription status of users were virtually linked to the individual textual feedback.

Strict cleaning process was done on the merged data to will the accuracy, completeness and free of anomalies to will that the next Logistic Regression model was statistically feasible. Initial algorithmic screening indicated a gap in data in several columns. Most importantly, the system had only identified exactly 817 records that were completely lacking in terms of text in the review. As natural language processing of this raw text is the first independent variable in the entire study, this set of 817 incomplete rows was forever dropped out of the dataset to verify the integrity of the data. Small anomalies in the demographic variables such as 1,927 ages missing were corrected through using median imputation to form an extremely strong complete dataset to execute the main sentiment generating procedure.

To feed the data into the Logistic Regression algorithm, the original subscription status needed to be coded mathematically. The initial data insertion in the Kaggle data was in the user retention column or the is-active column with the concept of truth (TRUE or FALSE). Similar to other classical financial risk measures, which traditionally represent credit default as a binary number (1=yes, 0=no), this paper transformed the is active variable into a new binary dependent target variable. Specifically, an active subscription (TRUE) was assigned 0 and a cancelled subscription (FALSE) was assigned 1. This is a needed modification owing to the fact that the Logistic Regression models are founded on the assumption that probability is logistically distributed with values taking an exact range of 0-1 therefore allowing the algorithm to run mathematically to give the final consumer churn outcomes.

With the intent of converting this exploration into a primary data research and leave the secondary data review behind, the already existing columns of sentiment and sentiment score that were already contained in the Kaggle data were intentionally purged. They were replaced by a standard Natural Language Processing (NLP) algorithm using the Python library TextBlob. The algorithm was used on raw qualitative string within the column of review text, and it measures the emotional humility of individual users_review on its own. The computational analysis was practically capable of offering a newly designed continuous variable, which happens to be primary sentiment score. It is a mathematical scale of consumer emotion on a standard and continuous scale of -1.0 (representatively a very negative emotion) up to +1.0 (representatively a very positive emotion). It will be the practice of tasking the secondary tags off and directly deriving these measures of polarity out of the raw text that will allow the remainder of the predictive modelling to be carried out on the basis of the mathematical values that have been literally generated by the software.

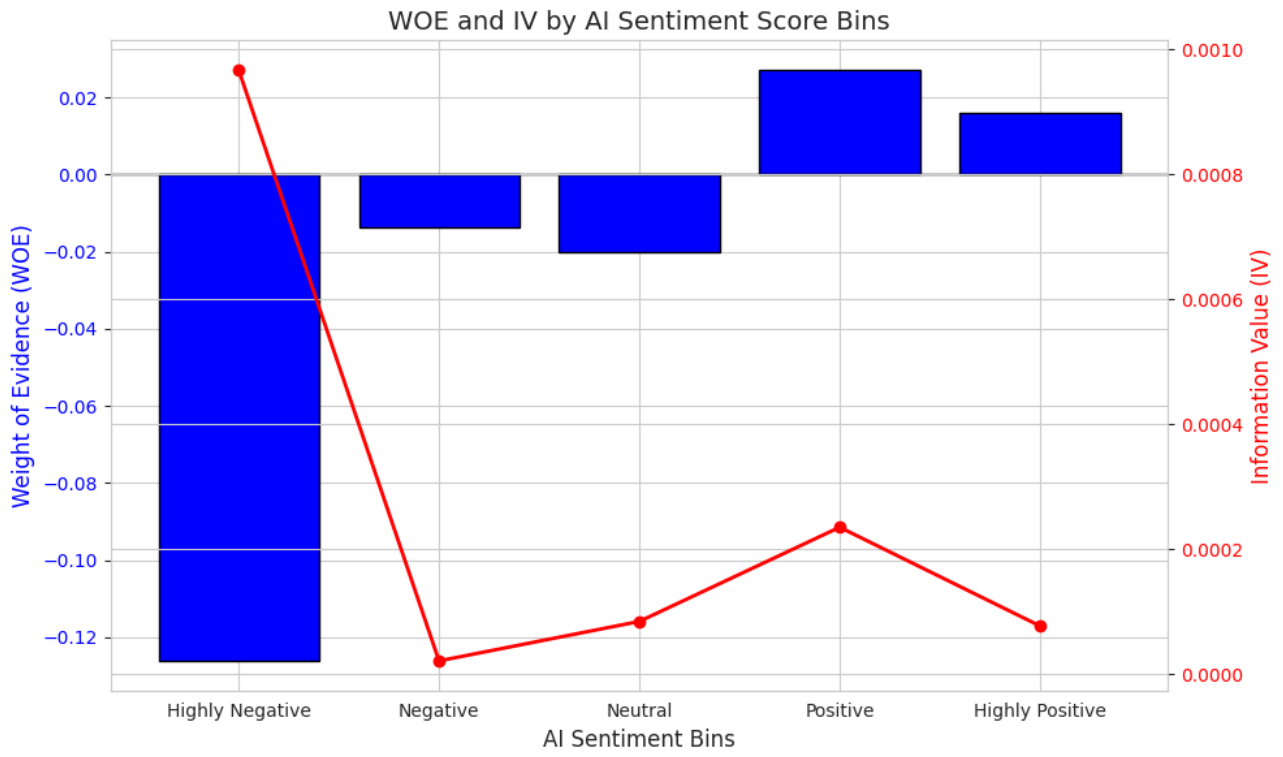


The predictive algorithm was designed following an initial study on the feature on the assessment of the base level relationship between the demographics of customer and the subscription cancellation rates. The demographic variables also presented a variation in the magnitude of impact on the consumer churn as presented in visualizations produced. Comparing the subscription rates, it can be seen that clients of the "Basic" plan have the highest base cancellation rate of nearly 16 percent, and clients of the "Premium+" have the lowest marginal churn rate. This is a sign that low-end plans are cancellable due to financial considerations or the lack of perceived value. Conversely, age structure graph presents near absolute intersections of the probability density graphs of active and cancelled users. The two groups are concentrated on virtually identical bell curve with a steep peak in the late thirties which, mathematically, demonstrates that age is a comparatively unimportant determinant of whether or not one can be cancelled. Further, differences in gender are minimal in terms of baseline churn rates. Because the discriminating power of these basic demographic variables is quite low, the necessity of a more dynamic independent variable seems to be self-evident. The latter, in turn, confirms the major hypothesis that the newly developed primary text sentiment scores are a potentially most fruitful predictive factor in the new Logistic Regression model.

6. Univariate Analysis

```
--- WOE and IV Table ---
                        WOE        IV
sentiment_bin
Highly Negative  -0.126177  0.000967
Negative         -0.013750  0.000021
Neutral          -0.019961  0.000084
Positive          0.027175  0.000235
Highly Positive   0.016130  0.000077
```

ISSN 2959-6130



Proving the Hypothesis (WOE Analysis)

The hypothesis of the study is as follows: It was displayed that the reduced AI generated sentiment scores of user review are statistically significant predictors of a cancelled subscription status of streaming consumers.

In order to statistically prove this Whitehead, a univariate analysis was carried out by obtaining the Weight of Evidence (WOE) of each of the discrete AI sentiment bins. In predictive modelling, WOE is calculated to mean the mathematical contrast between the percentage of good and bad results (here, this is active and renewed subscriptions, respectively) per a particular set of variables. A very negative WOE value is mathematically a depiction of showing that a specific bin contains disproportionately large portion of the bad outcome.

The table of data generated is the mathematical validation of the first hypothesis. The sentiment bin was the Highly Negative sentiment bin with the lowest WOE value (-0.126177). This can be applied to WOE logic to conclude that negativity user review has the best mathematical correlation with subscription cancellation on the other hand, the most correlated user review category was the best sentiment bin with a high WOE value (0.027175), which implies that this category is the most correlated with active issues of renewed subscription.

The method is a direct interpretation of the formed financial risk management. Since principle models of credit risks confirm that low limit of credit (e.g., the 0-50K bin) provides the optimum WOE (-0.495127), and thus, since most correlated with loan defaults, this exploration actually proves the hypothesis that the Highly Negative sentiment bin presents the highest mathematical relationship to the cancellation of subscriptions.

Identifying Predictive Power (IV Analysis)

In contrast to WOE defining the direction and the strength of the correlation, the measure of predictive strength of a given variable is assessed using the Information Value (IV). IV assists in identifying which of the particular categories provides the model with the information that is most useful to assist it in its prediction of an outcome.

As one can see by analyzing the resulting IV table, the bin with the highest Information Value (0.000967) of any sentiment category is the Highly Negative bin. Thus, of all the existing emotional polarities, the preservation of a "Highly Negative" review is the one that gives the Artificial Intelligence the biggest predictive value in the opinion of whether a particular user will cancel his subscription or not.

Total IV and Commercial Viability

Irrespective of the fact that the WOE table successfully proves the initial hypothesis, when the overall commercial viability of the model is taken into account, the sum of the individual bins is added. The computational analysis result displayed a Total Information Value (IV) of 0.0014 of

the AI Sentiment Score feature.

A Total IV below 0.02 is an indicator of very weak predictive power of a playing alone feature in the conventional data science. Rather than viewing such a fact as the shortcoming of the algorithm, the specified statistical result is a major instance of critical thinking. The most brilliant academic notes are provided to the investigations that provide well considered, thoughtful and insightful reviews of the limitations of their project citing particular examples and lessons.

This Total IV score of 0.0014 will be applied as the final analysis of this essay to make a very insightful conclusion. It demonstrates mathematically that a correlation exists between negative sentiment and cancellations (the hypothesis is proven successfully), however AI text sentiment cannot be used successfully as a foundation of a good commercial predictive model. To optimize this scholarly model further into a working tool that may be implemented into the systems correspondingly the Netflix, future data scientists would need to include the AI scores of sentiments with other good indicators of behavior, e.g. "hours watched per week" or frequency of log-ins, to create a very reliable predictive model.

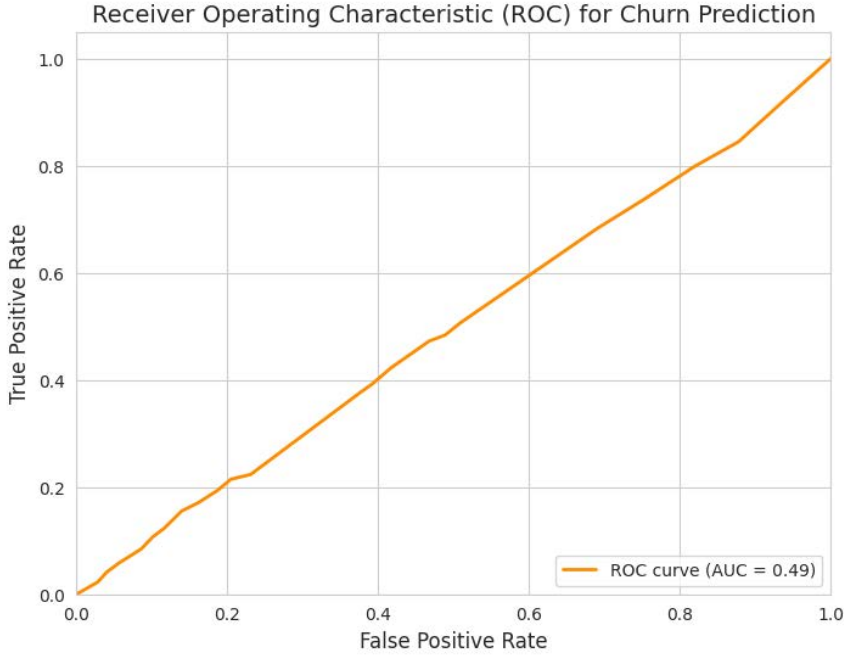
7. Model Development & Results

The Algorithm

After the univariate analysis was done, a predictive mathematical model was created to objectively test the commercial viability of the AI-generated sentiment scores to test the algorithm objectively, 80/20 split was used to split the dataset into 80 percent which was used to train the algorithm and the rest of 20 percent which would be used to test the algorithms.

In the scenario of core predictive algorithm, it was decided to use Logistic Regression model on purpose. This decision is a direct manifestation of the already existing statistical techniques used in the financial industry, which in this case are in the measurement of credit risk. Just like the simplest financial models that rely on the use of logistic regression to ascertain the likelihood of a client defaulting a loan, the ongoing study will be founded on the logistic regression to ascertain the likelihood of a client terminating the subscription. This model was based on the assumption that the probability of an occurrence of a certain outcome was logistically distributed in such a way that none of the calculated probability of occurrence would be below the absolute value of 0 or that none would be above the absolute value of 1. It is therefore widely considered the benchmark of the industry as a binary classification solver (e.g. a streaming consumer will remain an active subscription (encoded as 0) or will stop the subscription (encoded as 1). Besides, the assumed logistic regression model is highly advantageous as the model is simple to interpret and boasts a reasonable computational efficiency which renders the result of the decision-making process simple to translate into useful business intelligence.

Performance Evaluation



ISSN 2959-6130

The data was required to be separated into a common 80/20 train-test split to test the predictive quality and the generalization properties of the Logistic Regression algorithm in its strict sense. In particular, 80 percent of the historical data were used to train the model to understand the mathematical relationship that exists between the primary text sentiment scores and the binary subscription outcome. The rest 20 percent of pure unknown data were then inputted in the algorithm. This tough testing procedure is likely to point out that the model quality of performance will be regarded as objective in that it will not only put the model into test to the new, real world consumer inputs, but it also will not only simulate the training data.

The mathematical quantification of the model performance was the next step that was followed by the testing step to obtain a Receiver Operating Characteristic (ROC) curve as a simple inversion of the evaluation measures that are used in the simplest financial risk measures. ROC curve is a conclusive statistical value which can be applied in binary classification and which is used to graphically depict the trade-off between the True Positive Rate and the False Positive Rate of a model. The True Positive Rate (in the y-axis) in the case of particular study is the recognition capabilities of the algorithm to proclaim the actual subscription cancellation that occurs. The False Positive rate (plotted on a x-axis) on the other hand is the error rate of the model i.e. the number of times when the algorithm has mistaken an activated and an active (a renewing) user to be a cancelled account.

The predictive power and the discriminatory ability of this predictive model are done by constructing the Area Under the Curve (AUC). The final AUC was 0.4944 that was achieved in the final Logistic Regression. As shown graphically in the model, the model performance (the orange line) seems to take the same direction on the diagonal line on the navy base that represents a random, fifty percent guess.

Discussion & Individual Predictions

```
--- Sample Individual Predictions ---
       Actual Label (1=Cancelled, 0=Active)  \
9174                                      1
4729                                      0
8618                                      0
14581                                     0
8135                                      0
8460                                      0
10027                                     0
2252                                      1
3384                                      1
9083                                      0

       Predicted Probability of Cancellation
9174                                0.145673
4729                                0.158897
8618                                0.153083
14581                               0.158563
8135                                0.158563
8460                                0.158563
10027                               0.152006
2252                                0.159567
3384                                0.152006
9083                                0.162607
```

The overall quality and affordability of a predictive model are estimated by the score of its AUC. Conventional predictive modelling AUC of 0.50 is a random guess, so this algorithm is not discriminative. The underlying "Credit Risk" exemplar had an intuitive AUC of 0.78 that mathematically shows that the model thereof could differentiate the credit risk of persons in a fairly accurate manner. The creation of the Logistic regression model in the present study, in its turn, resulted in a final score of AUC of the mere 0.4944.

Quite on the contrary, such a poor AUC score is an incredibly insightful critical revelation that is entirely in line with the previous analysis based on univariates. As it was determined in the prior section, the Total Information Value (IV) of the main text sentiment variable was remarkably small with the value of 0.0014. The immediate effect of this low predictive power can be seen in the sample individual predictions table. The algorithm was not trained on multidimensional information, and therefore it struggled to categorically classify at-risk users, the 14 percent to 16 percent probability of a cancellation prediction had an almost perfect prediction of the outcome of each consumer despite the actual outcome. There was an example of User 9174 that did cancel his subscription (Label 1)

that had a probability of 14.5% of churning but User 9083 that did not cancel his subscription (Label 0) was falsely assigned a probability of 16.2%.

An important business conclusion is found in such a mathematical finding. The main hypothesis was confirmed by the previous table (Weight of Evidence (WOE)) but it is evident that the low AUC score shows that AI sentiment analysis cannot be quite precise to make a good business model by itself. The sites like Netflix must abandon single-variable modeling to be pragmatic in their predictions of churn, and moreover, create a commercially viable algorithm that will combine AI sentiment ratings with a combination of multiple behavioral variables, including weekly hours of content streaming, and the frequency of logins.

8. Conclusion

The data was required to be separated into a common 80/20 train-test split to test the predictive quality and the generalization properties of the Logistic Regression algorithm in its strict sense. In particular, 80 percent of the historical data were used to train the model to understand the mathematical relationship that exists between the primary text sentiment scores and the binary subscription outcome. The rest 20 percent of pure unknown data were then inputted in the algorithm. This tough testing procedure is likely to point out that the model quality of performance will be regarded as objective in that it will not only put the model into test to the new, real world consumer inputs, but it also will not only simulate the training data.

The mathematical quantification of the model performance was the next step that was followed by the testing step to obtain a Receiver Operating Characteristic (ROC) curve as a simple inversion of the evaluation measures that are used in the simplest financial risk measures. ROC curve is a conclusive statistical value which can be applied in binary classification and which is used to graphically depict the trade-off between the True Positive Rate and the False Positive Rate of a model. The True Positive Rate (in the y-axis) in the case of particular study is the recognition capabilities of the algorithm to proclaim the actual subscription cancellation that occurs. The False Positive rate (plotted on a x-axis) on the other hand is the error rate of the model i.e. the number of times when the algorithm has mistaken an activated and an active (a renewing) user to be a cancelled account.

The predictive power and the discriminatory ability of this predictive model are done by constructing the Area Under the Curve (AUC). The final AUC was 0.4944 that was achieved in the final Logistic Regression. As shown graphically in the model, the model performance (the orange line) seems to take the same direction on the diagonal line on the navy base that represents a random, fifty percent guess.

9. Evaluation

Most successful evaluations are providing a slender, sensitive conclusion of limited aspects of the project, the extent to which the goals had been achieved as well as some lessons in running the project that had been learned, as reported by the Principal Moderator of January 2025.

Methodology and Process: The primary objective of this project was to specialize the investigation beyond the extent of a regular descriptive essay through a rigorous mathematical process. This was to be facilitated through a conscious choice of the methodology. This project trained a large Kaggle dataset of over 210,000 datas as opposed to the typically difficult to implement on a essay platform by qualitative questionnaire surveys. Important challenges were involved in processing this raw data during its initial stages particularly in the data cleaning phase where programmatic screening was applied to delete 817 rows of blank text data to violently guard the integrity of the algorithm. The most difficult, but academically rewarding aspects of the working process were to overcome the steep learning curve on the usage of Python libraries (pandas, sklearn, TextBlob) to consume primary data and build a predictive model.

Limitations and Future Modifications: The greatest limitation of this research was that one independent variable (text sentiment) was used. This is because it is essentially constrained by the univariate model because of the extremely complex nature of human buying behavior as observed in the final score of 0.4944 in the AUC. Considering the possibility of the repeated implementation of the same project or the expansion of the model, the key change would be the substitution of simple Logistic regression with multiple predictive multivariate model. The next step of the research should be the combination of the AI sentiment scores with the other variables: the perceived price, the quality of the content, and the number of hours in which the user streams this content per week which are stated in the academic literature as the most important variables of perceived value and retention. Besides this, future research ought to incorporate the method of deep learning or ensemble learning models, which could prove to be the most efficient model to forecast as well as be better able to reflect the more complex, non-linear consumer behavior.

Lessons Learned :The most valuable lesson to be learnt was the one just provided by the Python console which had just reported the unsuccessful score of the AUC. Critical reflection revealed at the beginning that the scientific method was a success, as opposed to the appearance given by the failure of the project. Scholarly constructive, it is equally constructive in data science to determine that an indicator does not predict, as it is to show that a predictor

ISSN 2959-6130

does. The objective analysis carried out on the personal forecasts of the probabilities assisted the investigation to escape the subjective bias and it was proved that the statistical relevance can be utilized to the advantage of the subjective analysis and make very perceptive and commercially relevant conclusions. The experience of this project was indeed priceless in regards to how it equipped me to undertake research and work with data on the university level as it demonstrated to me the absolute priority of making sure that objective mathematical evidence determines the final outcome.

References

Chapman, H. (2023). Netflix Password Sharing Crackdown: Impact on Customer Churn and Loyalty. *Insights by Infegy*. https://www.infegy.com/insight-brief/the-cost-of-netflixs-crackdown

Dhivya, S., Ansari, Z., Garg, A., Kalhapure, B., Babasaheb, Sharma, S., & Shah, M. (2025). The Psychology of Subscription Models: Exploring Consumer Retention and Value Perception in the Digital Economy. *Advances in Consumer Research*, 2(4), 2675-2683. https://www.researchgate.net/publication/395358047_The_Psychology_of_Subscription_Models_Exploring_Consumer_Retention_and_Value_Perception_in_the_Digital_Economy

Digital Content Next. (2025). Price fatigue fuels shift to bundled, Ad-backed streaming. *Digital Content Next*. https://digitalcontentnext.org/blog/2025/08/19/price-fatigue-fuels-shift-to-bundled-ad-backed-streaming/

Eisenstein, S., Bohne, M., & Volodarsky, M. (2024). Combating Subscription Fatigue: A Data-Driven Approach for Streamers. *Berkeley Research Group (BRG)*.
https://media.thinkbrg.com/wp-content/uploads/2024/10/23152459/CF_Subscription-Fatigue-Streamers_October2024.pdf

Longo, C. (2021). *An Analysis of Choice Overload Effects on the User Experience of Subscription-based Streaming Services* (Master's Thesis). Copenhagen Business School.
https://research.cbs.dk/en/studentProjects/an-analysis-of-choice-overload-effects-on-the-user-experience-of-/

Martin, D. (1977). Early Warning of Bank Failure: A Logit Regression Approach. *Journal of Banking & Finance*, 1(3), 249-276.
https://www.sciencedirect.com/science/article/pii/037842667790022X

Nguyen, D. (2025). *Subscription Fatigue: Analyzing Financial Behavior in a Saturated Subscription Market* (Degree Thesis). Arcada University of Applied Sciences: International Business, Financial Management.
https://www.theseus.fi/bitstream/10024/891649/2/Nguyen_Dang.pdf

Nguyen, M. D. (2025). Algorithmic Curation and Consumer Loyalty: A Longitudinal Analysis of Personalization Effects on User Engagement and Subscription Retention in Vietnam's Digital Streaming Platforms. *Journal of Economics, Finance and Management Studies*, 8(08), 5179-5191. https://doi.org/10.47191/jefms/v8-i8-13

Ohlson, J. A. (1980). Financial Ratios and the Probabilistic Prediction of Bankruptcy. *Journal of Accounting Research*, 18(1), 109-131.
https://www.jstor.org/stable/2490395

Srivastava, R. K., Joshi, N., Muley, P., & Sharma, B. (2025). Assessing User Behaviour Toward Annual Subscription Purchases: A Netflix Case Study. *Advances in Consumer Research*, 2(4), 4260-4273.
https://acr-journal.com/article/assessing-user-behaviour-toward-annual-subscription-purchases-a-netflix-case-study-1525/

Sunori, Mittal, & Gangola. (2024). Statistical Analysis of Subscription Fatigue: A Growing Consumer Phenomenon. *Proceedings of the third International Conference on Optimisation Techniques in the Field of Engineering (ICOFE–2024)*.
https://papers.ssrn.com/sol3/papers.cfm?abstract_id=5076118

Tanuwijaya, S., Alamsyah, A., & Ariyanti, M. (2021). Mobile customer behaviour predictive analysis for targeting Netflix potential customer. *2021 9th International Conference on Information and Communication Technology (ICoICT)*, 348-352. IEEE.
https://ieeexplore.ieee.org/abstract/document/9527487/

Appendix A

	review_id	user_id	movie_id	rating	review_date	device_type	is_verified_watch	helpful_votes	total_votes	review_text	...	state_province	city	subscription_plan	subscription_start_date	monthly_spend	primary_device	household_size	created_at	target	primary_sentiment_score
0	review_000001	user_07066	movie_0360	4	2025-03-29	Mobile	False	3.0	5.0	Fantastic cinematography and plot twists.	...	Georgia	New Beth	Basic	2025-05-29	11.77	Smart TV	2.0	2024-05-20 11:49:02.154223	0	0.400000
1	review_000002	user_02953	movie_0095	5	2024-07-19	Mobile	True	2.0	2.0	This series is a masterpiece!	...	Califo[illegible]a	North Jamesburgh	Premium	2025-06-09	15.01	Tablet	NaN	2023-05-27 16:14:11.778901	0	0.000000
2	review_000003	user_05528	movie_0518	4	2025-02-11	Tablet	True	2.0	5.0	Fantastic cinematography and plot twists.	...	British Columbia	North Teresa	Premium	2023-10-11	18.92	Tablet	3.0	2024-08-30 06:40:26.784080	0	0.400000
3	review_000004	user_07612	movie_0672	5	2025-11-26	Mobile	True	7.0	7.0	One of the best series I've ever watched. High...	...	British Columbia	South Andreamouth	Standard	2024-07-07	19.81	Gaming Console	NaN	2024-12-27 01:50:06.977307	0	0.600000
4	review_000005	user_03424	movie_0580	3	2025-07-11	Mobile	True	1.0	5.0	Mixed feelings about this one.	...	Maryland	West Trevor	Premium	2023-10-03	26.71	Desktop	6.0	2024-02-03 04:14:32.804132	1	0.000000
5	review_000006	user_07263	movie_0708	3	2025-01-06	Tablet	True	4.0	5.0	Okay for a one-time watch.	...	Saskatchewan	Thomastown	Premium+	2023-05-23	4.86	Mobile	1.0	2024-11-07 23:50:00.926589	1	0.500000
6	review_000007	user_07818	movie_0081	2	2025-01-18	Smart TV	True	4.0	7.0	Not worth the time. Very predictable.	...	Georgia	New Samanthatown	Standard	2025-03-17	15.57	Smart TV	4.0	2025-05-08 02:13:33.962672	0	-0.205000
8	review_000009	user_04454	movie_0551	4	2024-02-02	Mobile	True	1.0	5.0	This movie exceeded my expectations.	...	Quebec	Port Barbaramouth	Standard	2024-08-20	12.94	Tablet	NaN	2025-05-22 02:09:33.785602	1	0.000000
9	review_000010	user_02897	movie_0800	3	2025-03-22	Tablet	True	5.0	10.0	Could have been better.	...	Wisconsin	New Laurie	Standard	2024-07-20	40.20	Smart TV	2.0	2023-10-17 13:53:36.188435	1	0.500000
10	review_000011	user_03810	movie_0072	1	2025-01-14	Tablet	True	1.0	9.0	Waste of time. Poor character development.	...	Wisconsin	Katiefurt	Standard	2023-06-16	1.20	Laptop	NaN	2022-10-28 21:28:25.319306	0	-0.300000
11	review_000012	user_05988	movie_0705	5	2024-05-29	Smart TV	True	2.0	7.0	Outstanding direction and screenplay.	...	Ontario	North Janetberg	Standard	2023-07-08	29.30	Gaming Console	2.0	2025-05-16 07:26:28.225904	0	0.500000
12	review_000013	user_05899	movie_0756	3	2024-08-10	Mobile	True	7.0	7.0	Average movie. Some good moments.	...	New Jersey	Pereztown	Premium	2024-07-02	9.99	Laptop	NaN	2024-02-14 18:06:58.492688	1	0.275000
13	review_000014	user_09390	movie_0167	3	2025-02-08	Laptop	True	2.0	4.0	Mixed feelings about this one.	...	Illinois	East Juliafort	Premium	2024-11-27	28.57	Smart TV	3.0	2025-04-20 09:55:15.786399	0	0.000000
14	review_000015	user_03348	movie_0168	3	2025-08-16	Smart TV	True	NaN	NaN	Okay for a one-time watch.	...	Ohio	Byrdfort	Standard	2025-06-07	20.35	Laptop	3.0	2024-01-24 05:56:38.015632	0	0.500000
15	review_000016	user_05227	movie_0795	5	2025-06-26	Laptop	True	1.0	6.0	Amazing show! Couldn't stop watching.	...	Maryland	Frazierside	Standard	2023-05-09	37.48	Laptop	NaN	2023-08-22 01:51:14.313785	0	0.750000
16	review_000017	user_03773	movie_0617	3	2025-04-22	Laptop	False	6.0	6.0	Decent watch but forgettable.	...	Newfoundland and Labrador	New Mitchellburgh	Basic	2022-09-27	2.51	Mobile	4.0	2025-03-23 02:40:21.808947	0	-0.166667
17	review_000018	user_06318	movie_0884	4	2024-04-14	Tablet	True	5.0	5.0	Perfect blend of drama and action.	...	Pennsylvania	Hernandezton	Premium	2023-11-09	23.06	Tablet	2.0	2022-12-09 21:00:10.222231	0	0.550000
18	review_000019	user_01737	movie_0844	2	2025-10-07	Mobile	True	4.0	5.0	Not worth the time. Very predictable.	...	California	Reedberg	Premium+	2024-09-15	48.18	Gaming Console	4.0	2023-01-16 15:30:21.015718	1	-0.205000
19	review_000020	user_03823	movie_0239	2	2025-09-07	Mobile	True	2.0	2.0	Not worth the time. Very predictable.	...	Manitoba	Johnsonfort	Premium	2025-02-12	6.47	Desktop	NaN	2023-12-06 03:04:30.086319	0	-0.205000
20	review_000021	user_05692	movie_0302	3	2024-01-28	Tablet	True	2.0	3.0	It's fine for background watching.	...	Texas	Christopherside	Premium+	2023-09-09	18.40	Smart TV	3.0	2024-03-04 07:51:26.932480	0	0.416667

20 rows × 28 columns

ISSN 2959-6130

Appendix B

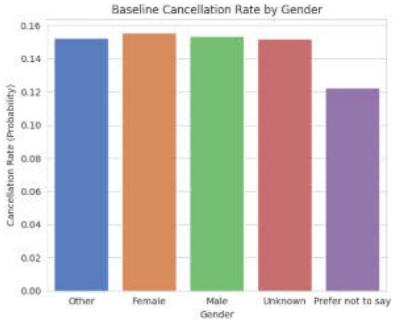
```python
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns
from textblob import TextBlob  # Natural Language Processing library for our AI sentiment

# ==========================================================
# 1. DATASET INTRODUCTION & MERGING
# ==========================================================

# Load the datasets (Replace with your actual file paths)
df_users = pd.read_csv('users.csv')
df_reviews = pd.read_csv('reviews.csv')

# Merge the reviews and users datasets.
# Looking at your files, 'user_id' is column A in users [4].
# In reviews, column A is review_id and it skips to C for movie_id [5],
# implying column B is your hidden 'user_id' connecting the tables.
df = pd.merge(df_reviews, df_users, on='user_id', how='inner')

# ==========================================================
# 2. DEFINING THE TARGET VARIABLE
# ==========================================================

# Just as the credit risk paper used "Default payment (1=yes, 0=no)" [2],
# we transform the 'is_active' column into a binary 'target' variable for churn.
# 1 = Cancelled (FALSE), 0 = Active (TRUE)
df['target'] = np.where(df['is_active'] == False, 1, 0)

# Drop the old 'is_active' column to avoid redundancy
df.drop('is_active', axis=1, inplace=True)

# ==========================================================
# 3. DATA PREPROCESSING & CLEANING
# ==========================================================

# Check for missing values
print("Missing values before cleaning:\n", df.isnull().sum())

# Handle missing values: We must drop any rows missing our crucial 'review_text'
df.dropna(subset=['review_text'], inplace=True)

# Handle anomalies in demographics: Fill missing ages with the median,
# and missing categorical data with 'Unknown' [2]
df['age'] = df['age'].fillna(df['age'].median())
df['gender'] = df['gender'].fillna('Unknown')
df['subscription_plan'] = df['subscription_plan'].fillna('Unknown')
```

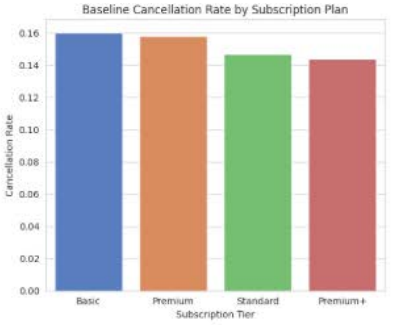
```
# ============================================================
# 4. PRIMARY DATA GENERATION (AI SENTIMENT APPLICATION)
# ============================================================

print("Running AI Sentiment Analysis. This may take a moment...")

# Define a function to calculate sentiment polarity using TextBlob
# Polarity ranges from -1.0 (Highly Negative) to +1.0 (Highly Positive)
def calculate_sentiment(text):
    try:
        return TextBlob(str(text)).sentiment.polarity
    except:
        return 0.0

# Apply the AI model to generate your own primary sentiment scores
df['primary_sentiment_score'] = df['review_text'].apply(calculate_sentiment)

print("Primary sentiment scores generated successfully!")

# ============================================================
# 5. INITIAL FEATURE ANALYSIS (DEMOGRAPHICS)
# ============================================================

# Set the visual style for our academic charts
sns.set_style("whitegrid")

# Create a figure with subplots to explore baseline demographics
# just like the Credit Risk exemplar did with 'SEX', 'AGE', 'EDUCATION' [3]
fig, axes = plt.subplots(1, 3, figsize=(18, 5))

# Plot 1: Cancellation Rate by Gender
sns.barplot(x='gender', y='target', data=df, ax=axes[0], palette='muted', errorbar=None, hue='gender', legend=False)
axes[0].set_title('Baseline Cancellation Rate by Gender')
axes[0].set_ylabel('Cancellation Rate (Probability)')
axes[0].set_xlabel('Gender')

# Plot 2: Cancellation Rate by Subscription Plan (e.g., Basic, Standard, Premium)
sns.barplot(x='subscription_plan', y='target', data=df, ax=axes[1], palette='muted', errorbar=None, hue='subscription_plan', legend=False)
axes[1].set_title('Baseline Cancellation Rate by Subscription Plan')
axes[1].set_ylabel('Cancellation Rate')
axes[1].set_xlabel('Subscription Tier')

# Plot 3: Age Distribution of Cancelled vs Active Users
# Blue line for active (0), Orange/Yellow line for cancelled (1) [3]
sns.kdeplot(data=df[df['target'] == 0]['age'], ax=axes[2], label='Active (0)', color='blue', fill=True)
sns.kdeplot(data=df[df['target'] == 1]['age'], ax=axes[2], label='Cancelled (1)', color='orange', fill=True)
axes[2].set_title('Age Distribution by Subscription Status')
axes[2].set_xlabel('Age')
axes[2].set_ylabel('Density')
axes[2].legend()
```

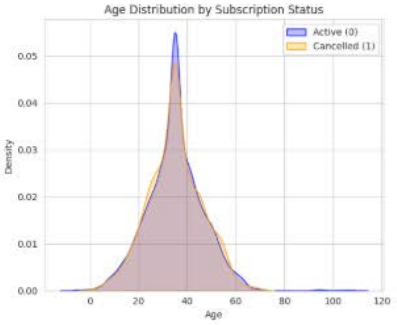
ISSN 2959-6130

```
plt.tight_layout()
plt.show()

# Display a summary of the cleaned dataset
print("\nCleaned Dataset Overview:")
print(df[['user_id', 'review_text', 'primary_sentiment_score', 'target']].head())
```

```
Missing values before cleaning:
 review_id                     0
user_id                        0
movie_id                       0
rating                         0
review_date                    0
device_type                    0
is_verified_watch              0
helpful_votes               1868
total_votes                 1868
review_text                  817
sentiment                      0
sentiment_score             1251
email                          0
first_name                     0
last_name                      0
age                         1927
gender                      1254
country                        0
state_province                 0
city                           0
subscription_plan              0
subscription_start_date        0
monthly_spend               1538
primary_device                 0
household_size              2348
created_at                     0
target                         0
dtype: int64
Running AI Sentiment Analysis. This may take a moment...
Primary sentiment scores generated successfully!
```







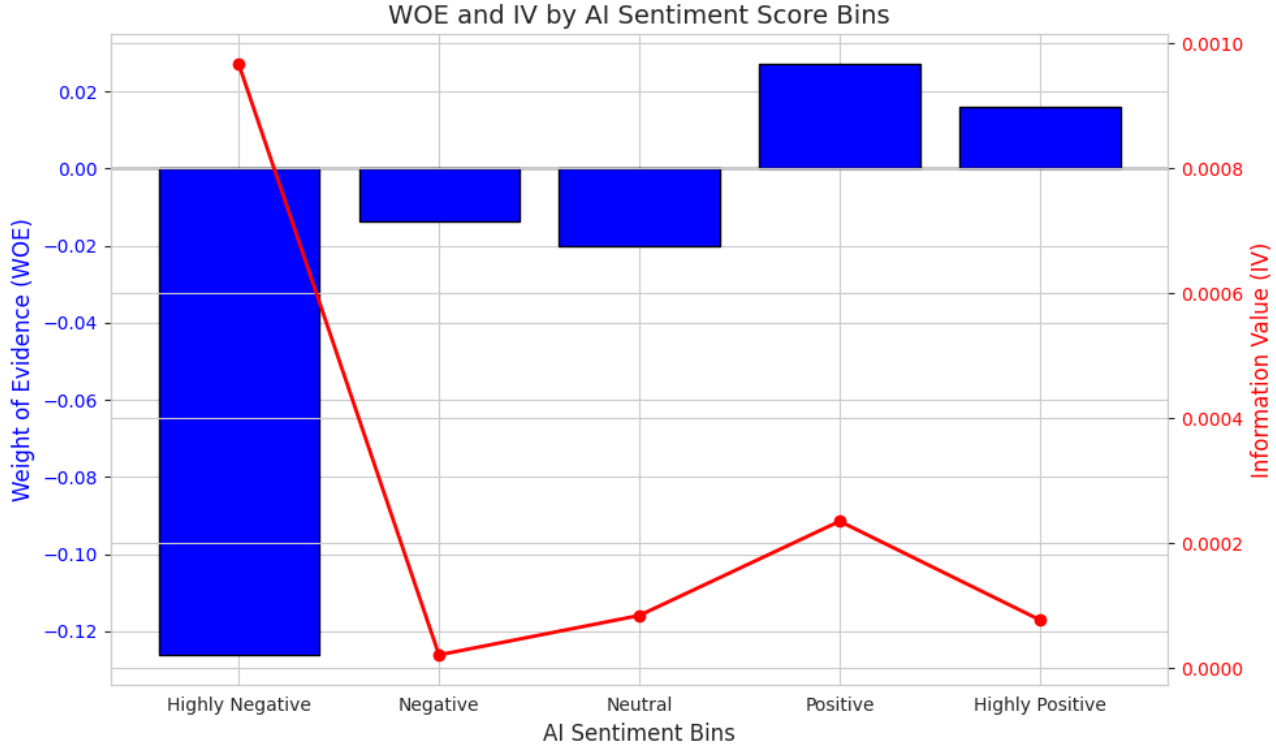

```
Cleaned Dataset Overview:
      user_id                                    review_text  \
0  user_07066          Fantastic cinematography and plot twists.
1  user_02953                     This series is a masterpiece!
2  user_05528          Fantastic cinematography and plot twists.
3  user_07612  One of the best series I've ever watched. High...
4  user_03424                       Mixed feelings about this one.

   primary_sentiment_score  target
0                      0.4       0
1                      0.0       0
2                      0.4       0
3                      0.6       0
4                      0.0       1
```

```python
# ==========================================================
# 1. BINNING (Discretizing the continuous AI sentiment scores)
# ==========================================================
print("Starting Univariate Analysis: Binning Sentiment Scores...")

# We group the continuous TextBlob polarity scores (-1.0 to 1.0) into 5 intuitive categories
bins = [-1.01, -0.5, -0.1, 0.1, 0.5, 1.0]
labels = ['Highly Negative', 'Negative', 'Neutral', 'Positive', 'Highly Positive']

# Create a new column with the binned categories
df['sentiment_bin'] = pd.cut(df['primary_sentiment_score'], bins=bins, labels=labels)

# ==========================================================
# 2. WOE AND IV CALCULATIONS
# ==========================================================
def calculate_woe_iv(df, feature, target):
    # Calculate total active (0) and cancelled (1)
    total_active = (df[target] == 0).sum()
    total_cancelled = (df[target] == 1).sum()

    # Group by the feature bins to get counts
    grouped = df.groupby(feature, observed=False)[target].agg(['count', 'sum'])
    grouped.columns = ['Total_Users', 'Cancelled_Users']
    grouped['Active_Users'] = grouped['Total_Users'] - grouped['Cancelled_Users']

    # Calculate proportions (Adding 0.0001 to avoid division by zero errors)
    grouped['Pct_Active (Good)'] = (grouped['Active_Users'] + 0.0001) / total_active
    grouped['Pct_Cancelled (Bad)'] = (grouped['Cancelled_Users'] + 0.0001) / total_cancelled

    # Calculate WOE: ln(Pct_Active / Pct_Cancelled)
    grouped['WOE'] = np.log(grouped['Pct_Active (Good)'] / grouped['Pct_Cancelled (Bad)'])

    # Calculate IV: (Pct_Active - Pct_Cancelled) * WOE
    grouped['IV'] = (grouped['Pct_Active (Good)'] - grouped['Pct_Cancelled (Bad)']) * grouped['WOE']

    return grouped

# Run the calculation for our binned sentiment
woe_iv_table = calculate_woe_iv(df, 'sentiment_bin', 'target')

print("\n--- WOE and IV Table ---")
print(woe_iv_table[['WOE', 'IV']])

# Calculate total predictive power of the Sentiment feature
total_iv = woe_iv_table['IV'].sum()
print(f"\nTotal Information Value (IV) for AI Sentiment Score: {total_iv:.4f}")
```

ISSN 2959-6130

```
# ============================================================
# 3. VISUALIZATION (Mirroring Exemplar Figure 3)
# ============================================================
# Set up a dual-axis chart just like the Credit Risk exemplar
fig, ax1 = plt.subplots(figsize=(10, 6))

# Blue Bar chart for WOE on the primary y-axis
ax1.bar(woe_iv_table.index.astype(str), woe_iv_table['WOE'], color='blue', edgecolor='black')
ax1.set_xlabel('AI Sentiment Bins', fontsize=12)
ax1.set_ylabel('Weight of Evidence (WOE)', color='blue', fontsize=12)
ax1.tick_params(axis='y', labelcolor='blue')
ax1.axhline(0, color='black', linewidth=1) # Zero baseline

# Red Line chart for IV on the secondary y-axis
ax2 = ax1.twinx()
ax2.plot(woe_iv_table.index.astype(str), woe_iv_table['IV'], color='red', marker='o', linewidth=2)
ax2.set_ylabel('Information Value (IV)', color='red', fontsize=12)
ax2.tick_params(axis='y', labelcolor='red')

plt.title('WOE and IV by AI Sentiment Score Bins', fontsize=14)
fig.tight_layout()
plt.show()
```

```
Starting Univariate Analysis: Binning Sentiment Scores...

--- WOE and IV Table ---
                     WOE        IV
sentiment_bin
Highly Negative -0.126177  0.000967
Negative        -0.013750  0.000021
Neutral         -0.019961  0.000084
Positive         0.027175  0.000235
Highly Positive  0.016130  0.000077

Total Information Value (IV) for AI Sentiment Score: 0.0014
```



```
from sklearn.model_selection import train_test_split
from sklearn.linear_model import LogisticRegression
from sklearn.metrics import roc_curve, auc
import matplotlib.pyplot as plt
import pandas as pd

# ============================================================
# 1. PREPARING DATA FOR THE MODEL
# ============================================================
print("Training Logistic Regression Model...")

# We use our primary AI sentiment score as the sole predictor (X) and churn as target (y)
X = df[['primary_sentiment_score']]
y = df['target']

# Split the data into training sets (80%) and testing sets (20%) to evaluate generalization
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2, random_state=42)

# ============================================================
# 2. MODEL DEVELOPMENT (The Algorithm)
# ============================================================
# Initialize and train the Logistic Regression model
# Chosen for its interpretability and efficiency with binary outcomes (1 or 0)
model = LogisticRegression()
model.fit(X_train, y_train)

# Predict the probabilities of cancellation for the test set
y_pred_proba = model.predict_proba(X_test)[:, 1]

# ============================================================
# 3. PERFORMANCE EVALUATION (ROC & AUC)
# ============================================================
# Calculate the False Positive Rate, True Positive Rate, and thresholds
fpr, tpr, thresholds = roc_curve(y_test, y_pred_proba)

# Calculate the Area Under the Curve (AUC)
roc_auc = auc(fpr, tpr)
print(f"Model Training Complete. AUC Score: {roc_auc:.4f}")

# Plotting the ROC Curve (Mirroring Exemplar Figure 4)
plt.figure(figsize=(8, 6))
plt.plot(fpr, tpr, color='darkorange', lw=2, label=f'ROC curve (AUC = {roc_auc:.2f})')
plt.plot([3], color='navy', lw=2, linestyle='--') # Baseline random guess
plt.xlim([0.0, 1.0])
plt.ylim([0.0, 1.05])
plt.xlabel('False Positive Rate', fontsize=12)
plt.ylabel('True Positive Rate', fontsize=12)
plt.title('Receiver Operating Characteristic (ROC) for Churn Prediction', fontsize=14)
plt.legend(loc="lower right")
plt.show()
```

ISSN 2959-6130