Leveraging Unstructured Call Log Data for Customer Churn Prediction

ARTICLE INFO

Keywords:
Customer Retention
Churn Prediction
Personality Mining
Call Log Analysis
Interpretable Machine Learning

ABSTRACT

Customer retention is important in the financial services industry. Machine learning has been incorporated into customer data analytics to predict client churn risks. Despite its success, existing approaches primarily use only structured data, e.g., demographics and account history. Data mining with unstructured data, e.g., customer interaction, can reveal more insights, which has not been adequately leveraged. In this research, we propose a customer churn prediction model utilizing the unstructured data, which is the spoken contents in phone communication. We collected a large-scale call center dataset with two million calls from more than two hundred thousand customers and conducted extensive experiments. The results show that our model can accurately predict the client churn risks and generate meaningful insights using interpretable machine learning with personality traits and customer segments. We discuss how these insights can help managers develop retention strategies customized for different customer segments.

1. Introduction

Customer relationship management (CRM) has always been a core business function for any company. Among the components of CRM, increasing customer engagement and loyalty is one of the most challenging tasks. Although customer acquisition and retention are both important, prior research [17] has showed that acquiring a new customer is typically five times more expensive than retaining an existing customer. Because of the high cost of customer acquisition, established businesses focus more on customer retention instead of acquisition. In customer retention, predicting customer churn risk is an important task. Each single percentage increase in customer churn prediction accuracy could potentially lead to a substantial revenue saving. This is particularly true for the financial services sector in which each customer may contribute to a considerable amount of profits, while customer engagement and loyalty are relatively low [42].

In this digital age, firms have been intensively relying on data analytics for customer churn forecast. With advances in machine learning, business intelligence applications can provide better customer insights by leveraging advanced data mining techniques powered by machine learning algorithms. Within the financial services field, a data-driven customer churn forecast model is essential for constructing efficient and effective retention strategies. Unfortunately, the lack of transparency and interpretability has limited the broader adoption of machine learning models and becomes a growing concern [1, 42]. In this research, we build an ensemble prediction model incorporating interpretable machine learning techniques to analyze multi-stacking data for churn forecast and customer retention, both applicable to financial services and other industries.

Churn decision of financial services customers can be associated with various factors, including customer demographics, behaviors, affective status, etc. The nature of churn prediction is a supervised learning problem that lies in feature construction and engineering, either from structured or unstructured customer data apart from cutting-edge learning algorithms. Existing features of churn prediction models are mainly derived from demographic data (gender, employment status, educational level, etc.), transactional data (investment decision, buying insurance, etc.), social network data (financial advisor, joint account, etc.), and other behavior data (risk-averse level, investment preferences, etc.). Most of these features are structured data in table format due to the ease of data collection directly from commercial databases. Structured data is often the common input for many off-the-shelf learning algorithms and knowledge-based systems. However, through empirical correlation analysis of such structured features with customer churn, we observed that the correlation coefficients are actually quite low, suggesting that using structured features standalone would not guarantee the satisfactory forecast performance (see Table 7. This motivates us to take features from unstructured data into consideration.

ORCID(s):

The cognitive process of churn decision making is complex and can be influenced by many indicators. For example, the affective status of customers has close indications to churn decisions which are reflected by various factors, ranging from products and services satisfactory level to personal feelings and opinions. Although basic communications data from customer service call center (e.g. frequency, call lengths, etc.) has been incorporated into customer churn prediction models in telecommunications [55], this interaction information is still used in a limited way (most often in form of structured features), which can only slightly improve the prediction accuracy. The content of the communication, which is unstructured data, has not yet been adequately utilized to capture finer granular customer insights for the financial gains of the firms. By deriving the customers' speaking pattern and personalities from unstructured data, we aim to predict churn risk more accurately and increase customer retention rate even further.

From the methodology aspect, researchers have investigated various kinds of feature engineering techniques with multiple extractions and stacking approaches. However, any further performance lift in terms of forecast accuracy using these approaches seems to be limited. Furthermore, those table-formatted features are usually accounted for only 20% of available customer data in a typical organization, while the majority 80% of the data obtained is the unstructured information [46]. Examples of such information include communication via phone calls, emails, messages, and social media channels. It is a much more challenging task to mine these types of customer data as they are often not obtained and stored on a regular basis.

In recent years, content mining combined with natural language processing (NLP) on unstructured data advanced the current predictive modeling, showing very promising results [48], e.g., opinion mining and sentimental analysis [39]. However, more finely granular customer affective statues than sentiment scores, which are carried by textual features, are not sufficiently exploited, indicating the research gap for customer churn prediction, which is the motivation for our research.

Our research seeks to answer three questions: 1) Is there any evidence that unstructured data can be helpful for customer churn prediction? 2) What are the approaches and techniques suitable to utilize this unstructured data for the churn prediction model? and 3) What are the profiles and characteristics of customers with high churn risk and how we can retain these clients? To answer these questions, we evaluate the effectiveness of unstructured data using different text mining methods. Benefited by the recent advancement in NLP and interpretable machine learning, we propose the churn forecast model that takes the benefit of these techniques to extract customer insights utilizing both structured and unstructured data. Particularly, we leverage the *textual features* extracted from customer call logs alongside other business data provided by our industry partner, a Superannuation management company, to enhance the customer churn prediction model. The churn forecast and interpreted machine learning model results serve as data-driven decision support that the firm would use to develop better customer retention strategies.

The contributions of our research can be broken down into four main points:

1. To the best of our knowledge, this is one of the first attempts to incorporate unstructured text mining with structured data mining and interpretable machine learning for churn prediction in the financial services.

2. We demonstrate that leveraging unstructured data and interpretable machine learning can capture a comprehensive customer preference spectrum for churn prediction. The model has been fully tested on large-scale real-world datasets containing two million calls from more than two hundred thousand customers.

3. We compare multiple text mining techniques to find suitable approaches and to fully leverage the unstructured data for churn prediction. We explore three textual feature representations and particularly personality traits in the research.

4. We are perhaps among the first to use interpretable machine learning in churn prediction. The proposed approach allows the evaluation of feature importance not only at the whole sample level but also at the customer segments and even individual customer levels. The insights from interpretable machine learning are useful in developing customized retention strategies for different cohorts.

The structure of this paper is organized as follows. Section 1 introduces the topic and current background in the financial services field with an emphasis on customer retention. In Section 2, we review the state-of-the-art literature on customer churn prediction focusing on recent text mining approaches as the motivation for our research work. Section 3 describes our model with technical details on unstructured data mining techniques and interpretable machine learning algorithms. In Section 4, we conduct extensive empirical experiments to evaluate the proposed approach using private

business datasets and discuss business insights from the interpretable machine learning analysis. Finally, Section 5 concludes our paper and highlights its potential for future work.

2. Literature Review

2.1. Customer Churn Forecast

Big data analytics has been playing a vital role in business information systems [20], which attracts numerous research efforts from both academia and industry sides, mainly focusing on customer analytics using big customer data [27]. A strategic decision support system for customer retention is the most important key to business success and has been extensively studied in recent years. Existing churn forecast approaches have utilized multiple machine learning algorithms to increase their prediction accuracy [5]. Almost all of these models are using only structured data [51]. Some of them have achieved great results with the tree-based algorithms and recent neural networks approach [22].

Different aspects of the business, have been tested for churn forecast, which revealed interesting insights on customer behaviors [44]. There are also various attempts [11] to incorporate customer call center data in such decision-support information systems [55]. However, the current state-of-the-art frameworks only focus on using the table-formatted data, e.g. the total number of calls, call duration [21]. This common practice is mainly due to the significantly high cost of obtaining and storing the unstructured data, e.g., the transcriptions of the calls, the text of the chats. Moreover, there are also other customer privacy and ethical concerns, which require further efforts and costs for the firms to obtain and anonymize the data. These are the main reasons why most current approaches do not consider unstructured data for their prediction models.

In the financial services field, client churn forecast has been extensively researched using different machine learning algorithms, particularly in the banking sector [4]. One of the most popular and best-performing algorithms is support vector machines, especially in the case of having an imbalanced dataset of credit card customers [15]. More advanced tree-based algorithms have been tested on electronic banking customer data [25] and achieved some positive results. A hybrid methodology combining k-Reverse-Nearest Neighborhood and One-Class Support Vector Machine (OCSVM) has been applied to solve similar problems [49]. Researchers have also attempted to combine fuzzy methodologies with machine learning algorithms [23] to increase prediction accuracy.

Considering the Superannuation industry in particular, customer retention strategies have been constructed with decision support systems using qualitative approaches, e.g., customer survey, focus group. Recently, big data analytics and feature engineering techniques have been applied in the Superannuation industry for customer retention [9]. The empirical results from their experiments show improvement in prediction accuracy. However, the results are not significantly different among multiple tested algorithms and show that these approaches cannot be further enhanced from the algorithm side using a similar type of structured data. We argue that the unstructured data which reveals other dimensions of consumer insights could be used to increase the prediction accuracy for better customer segmentation and retention strategies.

2.2. Big text analytics

Big text analytics has been proven to be effective in many business cases, especially in evaluating customer agility and engagement using online reviews [58]. Regarding the application of text mining in the financial services field, researchers have tested different techniques for sentiment analysis based on customer feedback and social media posts [57]. With the advancement of text mining research, there are various methods to be used to derive indicative user behavioral preference. A hybrid model consisting of concept-level sentiment and fuzzy formal concept analysis has been proposed to classify opinions from customer complaints [38].

In customer retention, researchers have taken into account the emotions based on text from customer emails to enhance churn prediction and achieve good results [12]. We believe there is richer lexical information carried by in the words and phrases spoken directly by customers instead of simple binary classification of positive or negative sentimental opinions. Recent research has tried to extract multiple term-based features as input for churn forecast models [48]. However, it is still unclear from these researches how textual information improves prediction accuracy in comparison with basic features and how it can help with retention strategies.

Table 1
Existing Text Mining Approaches in Customer Research

Paper	Features	Methods
Our paper	Basic Profiles, LIWC, TF-IDF, Word2Vec, Personality Traits	Multi-stacking Ensemble Prediction model with XgBoost, Logistic Regression, Gaussian Naive Bayes, Random Forest. Interpretable Machine Learning with SHAP-MRMR+.
[57]	Sentiment Scores, SentiStrength	K-Means clustering
[58]	TF-IDF	SVD-Based Semantic Keyword Similarity
[39]	TF-IDF, Sentiment Scores (Sentic Net3)	Fuzzy formal concept analysis and concept-level sentiment analysis
[12]	LIWC (only posemo and negemo)	Logistic Regression, SVM, Random Forest
[53]	LIWC, TF-IDF, Word2Vec	XgBoost
[7]	Big 5, Empowerment	Qualitative Survey, Multiple Regression Analysis
[3]	Big 5	Qualitative Survey, Statistical Analysis

2.3. Personality Mining

Personality mining has been researched by psychologists using mainly qualitative approaches [37]. Due to the high cost of data collection and annotation, personality datasets are relatively small in size and contain only written text data [32]. Current papers on personality mining mainly focus on feature extraction and engineering [43, 52], including neural network approaches using multimodal data [54]. The most common methods are Linguistic Inquiry and Word Count (LIWC) [36], Bag of Words, and other text sentiment analysis techniques.

On the other hand, personality mining has revealed significantly meaningful insights in terms of characterizing customer behavior, satisfaction [33] and loyalty in other consumer industries [26]. Researchers have proven that many personality traits are highly correlated with customer empowerment and satisfaction in the retail industry and suggest strategies based on these insights [7]. Customers in the financial services industry have similar characters to the retail consumers, yet there is little research on customer retention in this field using quantitative personality mining. Only qualitative survey methods have been applied to bank customers [3]. With the availability of spoken text data [38, 6], we can now apply the advanced transfer learning approach to the call logs with the similar spoken text nature in order to build a better churn prediction model which utilizes personality features.

2.4. Interpretable Machine Learning

In recent years, interpretable machine learning has been proposed to both explain the prediction model and extract business insights. Particularly, ensemble learning based on generalized additive models has been proposed to combine a traditional churn prediction model and explainable machine learning [13]. This method is mainly based on the improvement of the overall feature importance. On the other hand, Shapley Additive Explanations (SHAP) [30] value has been applied to related tree-based methods to explain both direction and magnitude of the features for every individual customer. Applying these methods in financial services can help reveal useful information on our different types of customers.

Realizing this research gap, we propose a more comprehensive churn prediction model leveraging both structured and unstructured big data and interpretable machine learning to extract meaningful insights and support managerial decisions. Table 1 presents a comparison of our work with related researches, the chosen text features and methods, highlighting the advantages and novelty of our model.

3. Methodology

The first part of our approach analyzes an integrated database of customer call log and profile data to construct churn prediction models. Our first hypothesis is that the multi-stacking prediction models using combined features can improve the churn prediction accuracy further than those using only the basic structured customer profile data.

We incorporate multiple text mining techniques on the customer call log datasets to extract four different feature sets: Term Importance, Phrase Embedding, Lexical Information, and Personality Traits. The rationale for using different text mining approaches is to capture insights about the customers at multiple granular levels, ranging from raw term frequency vector to underlying semantics space. Term Importance captures the lexical importance of bag-of-words from a given text corpus, e.g., sentence or paragraph. Phrase Embedding would help reveal the insights of the customer decision-making process based on the words used and their linkages to each other. Lexical Information would contain insights regarding latent concepts and topic-related terms used, which illustrates customer characteristics such as money-savvy or family-oriented. Finally, Personality Traits determine different customer psychological trait spectrum, e.g., Big Five personalities, which has a significant predictive power for the affective statuses and churn decision. After constructing the above textual features, we feed them into supervised prediction models alongside other features derived from structured data to predict customer churn probabilities. We also use interpretable machine learning with our SHAP-MRMR+ values to analyze the feature importance at three different levels: the whole sample, the customer segments, and individual customer levels.

3.1. Unstructured data mining

3.1.1. Term Importance

The most common technique to derive Term Importance features is the Bag-of-Words model. However, uni-gram or multi-gram models extract hundreds of thousands of textual features since we have almost one million call logs in our dataset. In our context, "term frequency-inverse document frequency" (TF-IDF) [44] is a more suitable technique as there are many common words with less insightful meaning, e.g. "hello" or "the". This methodology captures the major textual meaning by using TF-IDF expression. We derive almost 10,000 TF-IDF features in total. Table 2 gives a snapshot of the Term Importance matrix extracted from our call logs dataset.

Table 2
Sample TF-IDF features

Call ID\Term	"close"	"transfer"	"hello"	"yes"	"no"	Churn
Call 1	1.63	0.24	0.07	0.20	0.73	1
Call 2	0	2.19	0.07	0.27	0.15	1
Call 3	1.63	0.97	0.07	0.40	0	0
Call 4	0	0	0.07	0.27	0	0
Call 5	0	0.49	0.07	0.20	0.15	0
Call 6	0	0.24	0.07	0	0.15	0
Call 7	0	0	0	0	0.15	0

3.1.2. Phrase Embedding

Depending on the context, similar terms in English when combined in varied order can be understood differently. Considering the Term Importance as a single word feature alone would not be meaningful for our customer analysis. Therefore, we also aim to understand the semantics carried by various combination orders of terms by looking at the position and order of these phrases in a sentence, surrounding contexts, collocations, and their connections. To achieve this, we take into account the Phrase Embedding approach, which would help the forecast model to analyze the call logs transcript as sentences with contexts. We leveraged a widely used embedding algorithm, Word2Vec model [40] to extract a total of 50-word embedding features. Figure 1 illustrates an example of difference term relationships within the Word Embedding model which could help provide more meaningful insights for our prediction model.

3.1.3. Lexical Information

One of the most popular NLP researches is sentiment analysis, which classifies text into a binary dimension as positive or negative emotions. We believe there are more meaningful insights in other dimensions and topics of the English language. In order to extract this information, we leverage the Linguistic Inquiry and Word Count 2015 (LIWC) to extract latent concepts and topic-related text features. The LIWC 2015 dictionary contains about $6,400$ terms and emotions. Each term has a separate corresponding dictionary entry that identifies its one or more categories. For example, the term "disappointed" belongs to five different categories: "Verb", "Overall Affect", "Past Focus", "Negative Emotion" and "Sad". If the customer said the term "disappointed", all these five lexical categories' scores

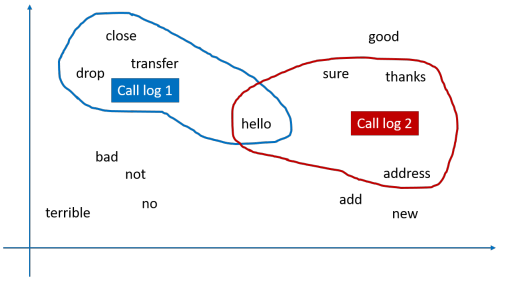


Figure 1: Word Embedding model captures relationships among terms. The terms are embedded in a feature space while maintaining the contextual relationships. In this example, the call log 1 shown in blue is a churned customer, and the call log 2 shown in red is a normal customer. Each term is represented by a vector, and the representation of a call log is obtained by joining the representations of all appeared terms.

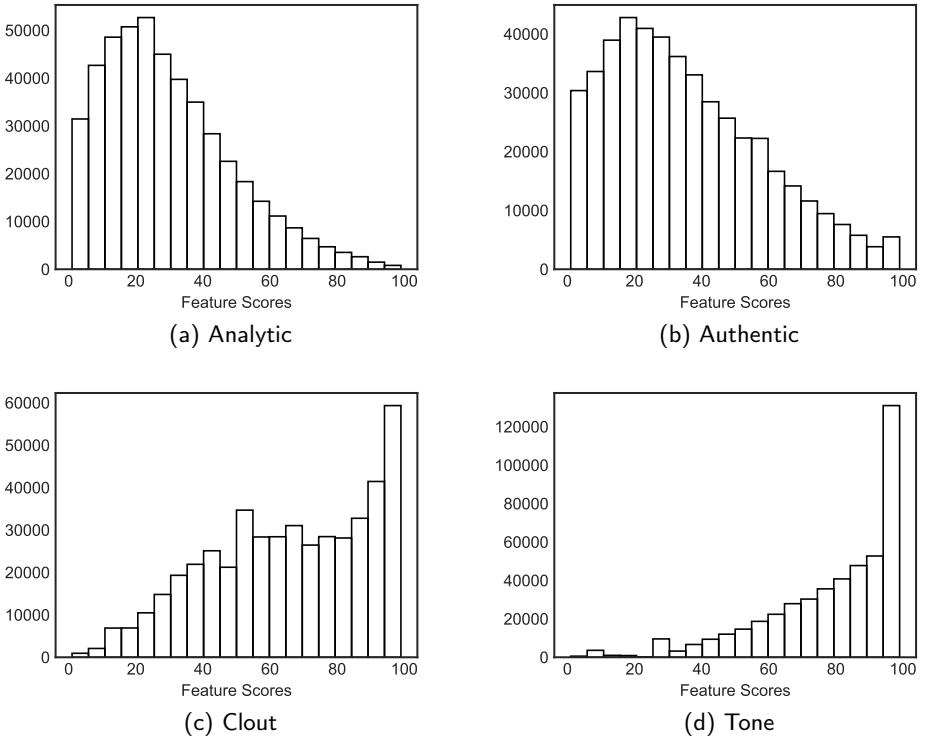


Figure 2: Histograms of LIWC 2015 main features

would increase respectively. This is one of the most comprehensive text mining dictionaries consisting of multiple topic-based categories ("money", "leisure"), emotions ("anger", "anxious") and speech-related features ("filler, "non-fluent"), which is more suitable for our call logs dataset. The LIWC 2015 has a total of 93 features. There are 12 punctuation-related features ("exclamation mark", "parenthesis", etc.). These features are only applicable for written text; hence they are omitted in our case of spoken text.

We also look particularly at features that can provide insights into customers' personalities. The histograms of some LIWC 2015 main features suggest that the majority of the customers are not very analytical and authentic (right-skewed "Analytic" and "Authentic"), but highly confident and emotional (left-skewed "Clout" and "Tone") in their speaking tone towards the call center agents (see Figure 2). Similar histograms can be seen in the sub-categorical features (e.g. "anger" and "anxious" features are nested under the category of the "Tone" feature). These features are extremely meaningful for us to distinguish customers and identify their individual personality traits.

3.1.4. Personality Traits

Personalities are human characteristics differentiated and reflected through their cognitive and behavioral patterns. Personality Mining is an advanced data mining technique to find the traits of a person from the way he or she speaks and acts. Psychology studies have shown cognitive language spoken as unique signals for different human behaviors. Individuals having distinguished traits often present themselves and behave differently. Knowing one's traits and understanding the differences in their preferences would help with communicating and connecting to the person on a more personalized level. Our personality mining model incorporates the Five-Factor Model of Personality (Big Five) [19] for the personality mining task. The Big Five model contains five fundamental human traits: openness to experience, conscientiousness, extroversion, agreeableness, and neuroticism (Table 3 presents the typical characteristics for each of the five traits). These traits, widely accepted by psychologists as a standard measure, are proven to stay consistent despite age, gender or cultural background [10].

Table 3
The Big Five Personality Traits

Trait	High Rank	Low Rank
Extroversion	Outgoing, active, seek excitement	Aloof, quiet, enjoy time alone
Neuroticism	Prone to stress, negative emotions	Emotionally stable, self-satisfied
Conscientiousness	Organized, punctual, hard-working	Spontaneous, careless, hedonistic
Agreeableness	Trusting, empathetic, compliant	Uncooperative, not listen to others
Openness	Creative, imaginative, curious	Practical, conventional, skeptical

The rationale of using personality traits in churn prediction is that customers with varied personalities might make different financial decisions. From the model perspective, personality mining could be treated as a supervised learning task, and the results can explain customer behavior sufficiently. In this research, we use a random forest algorithm to train the prediction model for personality based on collected training data, e.g., the First Impression dataset. Then the trained model is transferred for use in identifying the Big Five traits of the customer based on their call logs.

3.2. Multi-stacking Ensemble Model for Churn Prediction

The customer churn forecast model automatically extracts all mentioned textual features, then combines with customer profile data from various business databases as input for the multi-stacking churn forecast models. The prediction model computes a ranked list of customers with various churn scores, upon which the company can decide which customers are at a higher risk to churn.

We apply a multi-stacking ensemble method to build our final prediction model, as illustrated in Figure 3, by ensemble the predicted churn risks from single stacking models using only one type of features. This approach reduces the model training time significantly compared to the model with all features being linearly combined. The required time for training on Pension and Investment datasets are cut down to half from approximately 45 minutes to 23 minutes, while the computation speeds of the other two bigger dataset models are almost three times faster (only took about 31 minutes instead of 86 minutes). The predicted churn risk is computed as:

$$\hat{Y} = f_7(Y) = f_7(Y_{CP} + Y_{Text}) = f_7(Y_{CP}) + f_6(Y_{PT} + Y_{LI} + Y_{PE} + Y_{TI}) \tag{1}$$

$$\iff \hat{Y} = f_7(f_5(X5) + f_6(f_4(X4) + f_3(X3) + f_2(X2) + f_1(X1))) \tag{2}$$

where Y_{CP}, Y_{TI}, Y_{PE}, Y_{LI}, and Y_{PT} represent the predicted churn risks using the single feature set of Customer Profiles, Term Importance, Phrase Embedding, Lexical Information, and Personality Traits accordingly. The structure of the model can be viewed as a multi-layer network where the output of the previous layer is used as the input for the next layer (see Figure 3).

In our multi-stacking ensemble model, we chose four supervised learning algorithms to build the churn forecast models, namely Naïve Bayes, Logistic Regression, Random Forest, and Extreme Gradient Boosting. They represent linear models and state-of-the-art ensemble models. We briefly describe their ideas as follows:

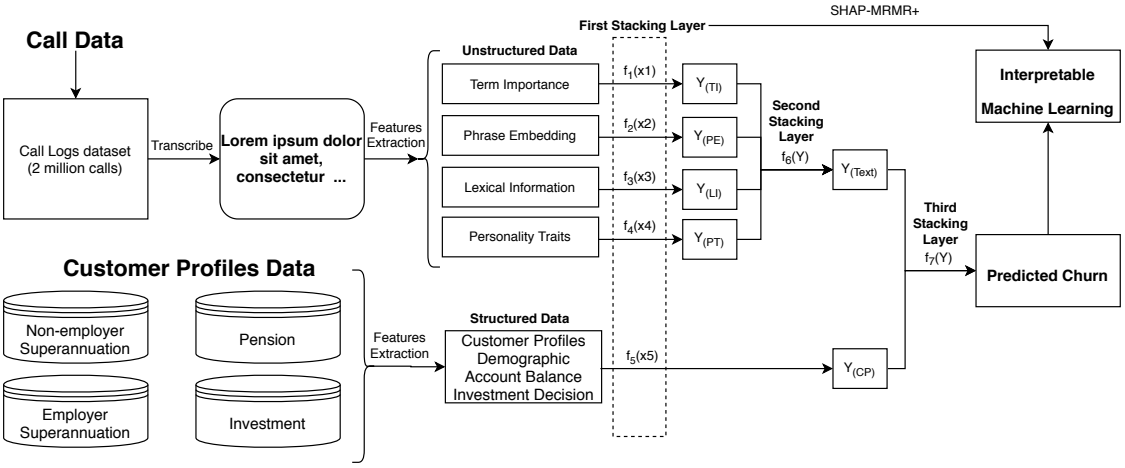


Figure 3: Multi-stacking Ensemble Model and Interpretable Machine Learning

3.2.1. Gaussian Naïve Bayes (NB) Classifier

This supervised learning algorithm is built upon the famous Bayes' theorem. The computational foundation is based on the "naïve" assumption that all pairwise features are not correlated with each other. We denote the churn label as y and the dependent feature vector as x_1 through x_n where all features are pairwise uncorrelated. Under naïve independence assumption, the following classification rule can be used:

$$P(y \mid x_1, ..., x_n) \propto Py \prod_{i=1}^{n} P(x_i \mid y) \tag{3}$$

$$\implies \hat{y} = \arg\max_{y} P(y) \prod_{i=1}^{n} P(x_i \mid y), \tag{4}$$

where $P(x_i \mid y)$ can be obtained via Maximum A Posteriori (MAP) estimation.

3.2.2. Logistic Regression (LR) Classifier

Logistic Regression is derived from the statistical method of analyzing one or more uncorrelated variables. The customer churn decision is evaluated as a dichotomous variable (in which there are only two possible outcomes of churn or not churn). The algorithm's target is to optimize the best fitting model to measure the correlation between customer churn decision and all other textual and standard account features. By implementing a L2 regularization, the binary class L2 penalized algorithm optimizes the corresponding cost function $\boldsymbol{J}$:

$$\boldsymbol{J} = \min_{w,c} \frac{1}{2} w^T w + C \sum_{i=1}^{n} \log\left(\exp\left(-y_i(x_i^T w + c)\right) + 1\right) \tag{5}$$

where w represents the weight, c represents the cost, x_i and y_i represent the feature set and churn label of customer i respectively. Other setup and parameter tuning are the same as in the baseline model of similar state-of-art work [9].

3.2.3. Random Forest (RF) Classifier

A Random Forest Classifier is a classifier based on a random forest family of classifiers based on a family of classifiers $h(x \mid \Theta_1), ..., h(x \mid \Theta_K)$ based on a classification tree with parameters Θ_k randomly chosen from a model random vector Θ. For the final classification $f(x)$ which combines the classifiers $\{h_k(x)\}$, each tree casts a vote for the most popular class at input x, and the class with the most votes wins. Specifically given data $D = \{(x_i, y_i)\}_{i=1}^{n}$, we train a family of classifiers $h_k(x)$. In our case, each classifier $h_k(x) \equiv h(x \mid \Theta_k)$ is a predictor of n and $y = \pm 1$ is the outcome associated with input x.

3.2.4. Extreme Gradient Boosting (XGB) Classifier

The algorithm was introduced by [8] in 2014 as an enhanced version of the greedy gradient boosting machine [18]. XgBoost has become one of the most widely-used and effective algorithms in supervised machine learning. In our model, we first define the tree $f(x)$ as:

$$f_t(x) = w_{q(x)}, w \in R^T, q : R^D \rightarrow \{1, 2, ..., T\} \tag{6}$$

where w represents the vector of the scores on the leaves, q is a data assigning function for the corresponding leaf, and T represents the total number of defined leaves. We also define the gradient g_i and the Hessian (second-order derivative) h_i as follow:

$$g_i = \partial_{\hat{y}_i^{(t-1)}} l\left(y_i, \hat{y}_i^{(t-1)}\right) \tag{7}$$

$$h_i = \partial^2_{\hat{y}_i^{(t-1)}} l\left(y_i, \hat{y}_i^{(t-1)}\right) \tag{8}$$

where y_i represents the churn label and $\hat{y}_i^{(t-1)}$ represents the predicted churn at time $(t-1)$. Using regularization to improve the generalization performance, the objective function at t^{th} tree is:

$$obj^{(t)} = \sum_{i=1}^{n} \left[g_i w_{q(x_i)} + \frac{1}{2} h_i w^2_{q(x_i)}\right] + \gamma T + \frac{1}{2}\lambda \sum_{j=1}^{T} w_j^2 + \alpha \sum_{j=1}^{T} \left|w_j\right| \tag{9}$$

where w_j represents the weight assigned to the j^{th} leaf, $\gamma = 0$, $\alpha = 0$ and $\lambda = 1$ are predefined parameters for the penalization and regularization terms respectively.

3.2.5. Bidirectional Encoder Representations from Transformers (BERT)

We also consider another state-of-art text mining approach using the Bidirectional Encoder Representations from Transformers (BERT) embedding [14] and Bidirectional Long Short-Term Memory (BiLSTM) neural network [45]. For BERT, we utilized the pre-trained model Transformer BERT [41] to encode our call logs with a standard 768 embedding size, which is significantly larger than the 50 embedding size of our Word2Vec model. We built a fully-connected BiLSTM neural network using Tensorflow Keras with two BiLSTM layers and one dense layer as in Table 4. The model was trained for 10 epochs with 'relu' activation and 'Adam' optimizer.

Table 4
Fully-connected BiLSTM neural network with BERT Embedding Model

Layer type	Output Shape	Parameters
BiLSTM Layer 1	(None, 768, 128)	33792
BiLSTM Layer 2	(None, 64)	41216
Dense Layer	(None, 1)	65
Total parameters: 75,073		
Trainable parameters: 75,073		

3.3. Interpretable Machine Learning for CRM

3.3.1. SHAP-MRMR+

SHAP [30] is a unified model to interpret many machine learning models, which connects game theory with local explanations. It combines multiple methods into one consistent and locally accurate additive feature attribution with the expectation-based approach. Given a prediction $f(x)$ and $S \subseteq Z/\{i\}$ where Z is the set of all input features and M is the number of "interpretable" inputs, we calculate the Shapley values [31] as a weighted sum of the impact of each feature i added to the model, which is averaged over all possible orders of features:

$$\phi_i(f, x) = \sum_{S \subseteq Z/\{i\}} \frac{|S|!\,(M - S - 1)!}{M!} [f_x(S \cup \{i\} - f_x(S)] \tag{10}$$

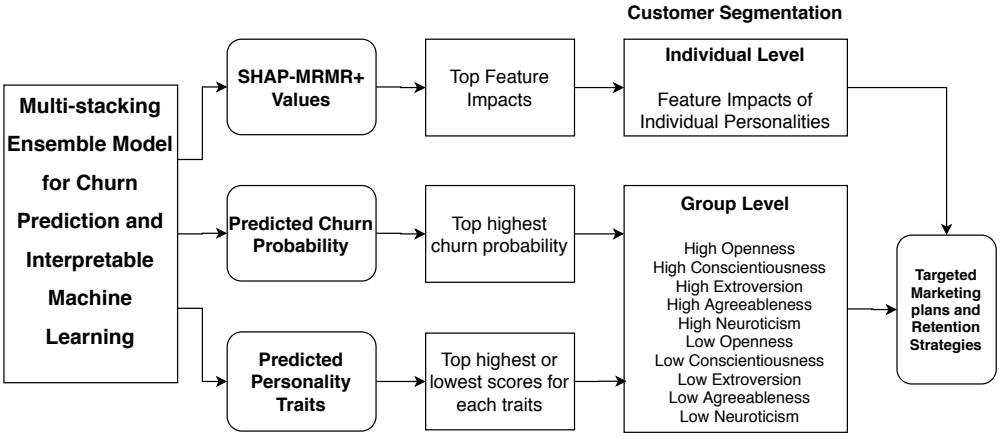


Figure 4: Method for CRM strategies based on Personalities and Interpretable Machine Learning

In this paper, we propose a modified SHAP to interpret our machine learning model. Minimum Redundancy Maximum Relevance (mRMR) [35] is considered more powerful than the maximum relevance feature selection. It can select features that are mutually far away from each other but still have a "high" correlation to the classification variables. We combine the Shapley value as in th equation above and the positive mRMR value to compute the SHAP-MRMR+ as:

$$\text{SHAP} - \text{MRMR}_i(f,x) \ = \frac{1}{2}\left(\phi_i\,(f,x) + \sum_{i=1}^{N} \frac{|\text{MRMR}_i| + \text{MRMR}_i}{2N}\right) \tag{11}$$

where MRMR_k is the computed mRMR value. Our intuition is that the positive mRMR values indicate the most important features globally, which impacts might be minimized under the local Shapley calculation. By adding these values to the SHAP values, we can further highlight these features to support managerial decisions both globally and locally for every individual customer.

3.3.2. Customer segmentation with Personality

Our proposed model leverages the predicted customers' personalities and churn risks to construct suitable targeted marketing plans. First of all, we apply a multi-filtering technique to identify the high-value customers with the highest likelihood to leave the company. We segment this customer database into ten different categories based on whether they have high or low scores in each of the Big Five personality. We then combine with the marketing database to develop different retention strategies for each group. For example, if the customers are in the "High Openness" group, direct email marketing can be a cost-effective strategy to keep them stay with the company. On the opposite side, for customers in the "Low Openness" group, our model suggests other approaches such as promotional campaigns to engage them more. By segmenting customers and proposing different churn prevention strategies, we can improve both the effectiveness and efficiency of the CRM plans in terms of timing, cost, and resources. Figure 4 illustrates the process of generating the managerial implications with combined customer segmentation using Personality Traits and interpretable machine learning with SHAP-MRMR+.

4. Empirical Experiments and Results

4.1. Datasets

Customer Call Logs dataset (CL) : We obtained the call center data directly from an Australian financial services company. The firm has recorded over three million customer calls to the hot-line for the period between 04/2011 and 04/2017. The company also outsourced the text transcriptions of the calls to a third-party service provider. For each call, conversational dialogues are separated into two monologues of the call center agent and the customer respectively. Almost two million calls can be identified using unique customer IDs for mapping back to the existing client database. Additionally, some customers call to end the services and specifically use related terms such as "close account" or

"terminate account". We exclude these calls recorded within 14 days range of an account closing request by customers to reduce bias prediction and false positive prediction. Our model uses the text transcriptions of only the identified customer calls to proceed further. The final CL dataset has also been processed to remove all private names, account identification terms, and sensitive financial information to ensure anonymity and data protection for the customers. Each call in our dataset contains on average 314 words in length, with the median word count is 240 words.

First Impressions dataset (FI): For Personality Traits, we use a public dataset called First Impression [38] to train a personality prediction model and then transfer learning on our private CL datasets. Though there are multiple public datasets (see Table 5), we focus on FI due to its nature of being spoken text and large sample size, which is suitable for our context. The dataset contains ten thousand videos of people speaking directly to the camera. The mean duration of each video is about 15 seconds, 43 words spoken per clip on average. 435, 984 words (183, 861 non-stop words) has been transcribed for all ten thousand clips. Psychology experts have labeled each video using the scoring system ranged [0,1] from the Big Five personality model.

Table 5
Comparing Personality Datasets

Dataset	Sample Size	Description
First Impression [38]	10000	Spoken text at varied length on multiple topics
Youtube [6]	404	Spoken text at varied length on multiple topics
Personage [32]	320	Short spoken text on restaurant topic
myPersonality [28]	154000	Short written text on multiple topics (Facebook statuses)
Essays [37]	2468	Long written text on multiple topics

Customer Profiles dataset (CP) : We have in total four different Customer Profiles databases provided by an Australian financial services company. The first one is the Employer superannuation dataset containing institutional customers who are employers of other employee accounts. Therefore, it has more features related to the demographic data and sub-account information. We have 133 features in this dataset in comparison to 106 features for other datasets. Non-employer superannuation dataset contains individual customers who open the accounts without reference from their employers. The Pension dataset includes individual superannuation customers who have reached their retirement age or opted for early retirement and their superannuation accounts become pension accounts. The three datasets are quite similar in terms of financial behaviors. The investment dataset is a little bit different with more features related to investment decisions. The distinguishable characteristics of the four datasets provide us with a better experimental set up to prove the generalizability of our methodology, that is, it would work for different types of data and customers not only in the financial services sector but also for other industry as well.

Table 6 presents the sizes and basic statistics of the four databases. They include several demographic features (e.g., sex, age, address) and financial performance features (e.g. annual investment rate, total balance). Combining with the CL dataset, the merged datasets in Table 4 consist of only customers who have made phone calls to the company hotlines since 2014. The churn decisions of these customers have been labeled as 1 for churn and 0 for not churn. The binary indicator serves as the ground truth Y for testing and evaluating our churn forecast model.

Table 6
Statistics of the merged datasets

Datasets	Number of Features	Total	Churn	Churn (%)
Non-employer Super	106	49,996	2,067	4.13%
Employer Super	133	36,608	1,959	5.35%
Pension	106	28,046	582	2.08%
Investment	106	23,422	2,098	8.96%

We also perform correlation analysis on all CP basic features using the Pearson's r [34], the Spearman's ρ [47] and the Kendall's τ [24] correlations. Since we have a various combination set of basis features, ranging from numeric to categorical variables with different distributions, using multiple metrics would provide an unbiased analysis.

The result in Table 7 presents the top five correlated features using each analysis strategy. As we can see, these features are account balances or variables related to the outflow of investment (e.g. outflow recency, outflow frequency, outflow amount, outflow ratio, etc.). This general financial behavior is predictable with common features in the financial services industry. However, the correlation coefficients of all the basic features are really low. The scores range is from −0.085 to 0.065, while meaningful correlated features should have higher scores than 0.5 or lower than −0.5. Therefore, we believe the incorporation of text features would help the prediction accuracy even further.

Table 7
Correlation analysis of basic features and customer churn label

Pearson's r		Kendall's tau		Spearman's rho	
Features	Scores	Features	Scores	Features	Scores
Outflow recency	0.0654	Outflow recency	0.0609	Account balance	−0.0851
Call recency	0.0481	Account Balance	−0.0695	Outflow recency	0.0649
Account growth	0.0471	Outflow frequency	−0.0579	Outflow frequency	−0.0607
Number of options	−0.0436	Outflow amount	−0.0552	Outflow amount	−0.0590
Saving plan N	0.0430	Outflow ratio	−0.0517	Outflow Ratio	−0.0552

4.2. Baselines and Evaluation Metrics

To evaluate the performance of our models, we built four baselines, which are the churn prediction models using only the basic features of the four different customer datasets. Regarding evaluation metrics in our case, the company is specifically interested in identifying customers with the highest churn risks. They intend to target at the top 30% of clients with high churn risk. Therefore, we build our model to predict churn risk instead of binary classification. Our predicted churn p is ranging from 0 to 1 as the probability for churn. The experiments run with 10-fold cross-validation and the performance results are averaged for all folds. We use Area-Under-the-Curve (AUC) scores as our evaluation metric:

$$AUC = \frac{1}{N_1 N_0} \sum_{i=1}^{N_1} \sum_{j=1}^{N_0} \mathbf{1}_{(p_i > p_j)} \tag{12}$$

where N_1 represents the total number of churn customers (true label 1), N_0 represents the total number of not churn customers (true label 0), p_i and p_j represents the probability scores assigned by our model to each label respectively. $\mathbf{1}$ is the indicator function with value equals 1 if $p_i > p_j$ and 0 otherwise.

4.3. Churn Prediction Results

First of all, we designed a preliminary experiment to test the correlation between the word embedding size and the AUC scores using 10-fold cross-validation with the Logistic Regression algorithm on the Investment dataset. The result in Figure 5 shows that the prediction accuracy did not improve significantly when we increase the size over 50. Therefore, we only use 50 Word2Vec features in our prediction model.

We test the text features individually to confirm the effectiveness of unstructured data as input for churn prediction. We build separate models with XGB algorithm using Term Importance (TI), Phrase Embedding (PE), Lexical Information (LI), and Personality Traits (PT) text features. The result in Figure 6 shows that multiple text mining techniques can capture different information. It is also noticed that models with TI and LI features outperformed others, indicating the simpler text mining techniques might yield a sufficient result already. In the case of PT, the model can achieve comparable results even with only five features. This shows that personalities are good indicators for evaluating client churn risk. Furthermore, we believe by stacking all approaches together, the features complement each other to achieve an even better result in the final churn prediction model with AUC score 0.8124 (see Table 8).

To evaluate the performance of text mining techniques, we test separate churn forecast models using different feature sets: (1) using only structured data from the Customer Profiles (CP) datasets, (2) using only unstructured data from the Call Logs (CL) datasets, (3) using the multi-stacking ensemble feature set to build a churn forecast model.

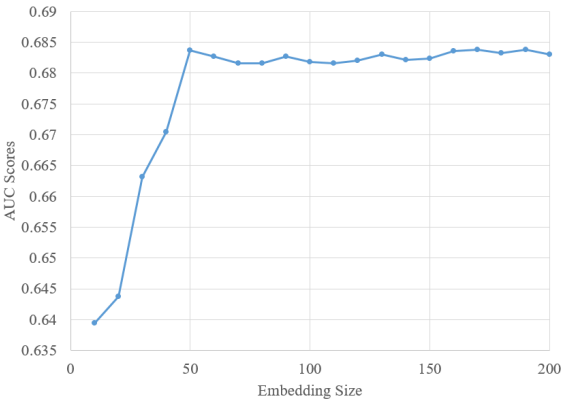


Figure 5: Correlation between Word Embedding Size and AUC Scores

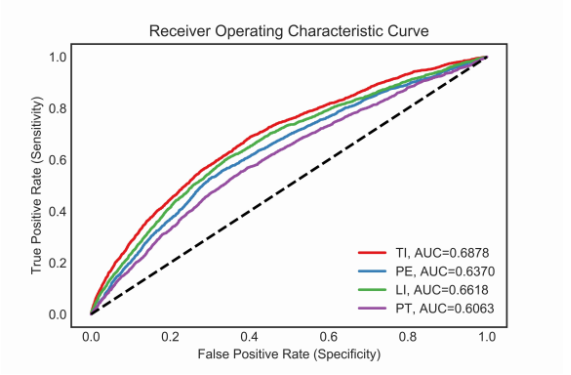


Figure 6: Text Features Churn Prediction Model for Investment dataset

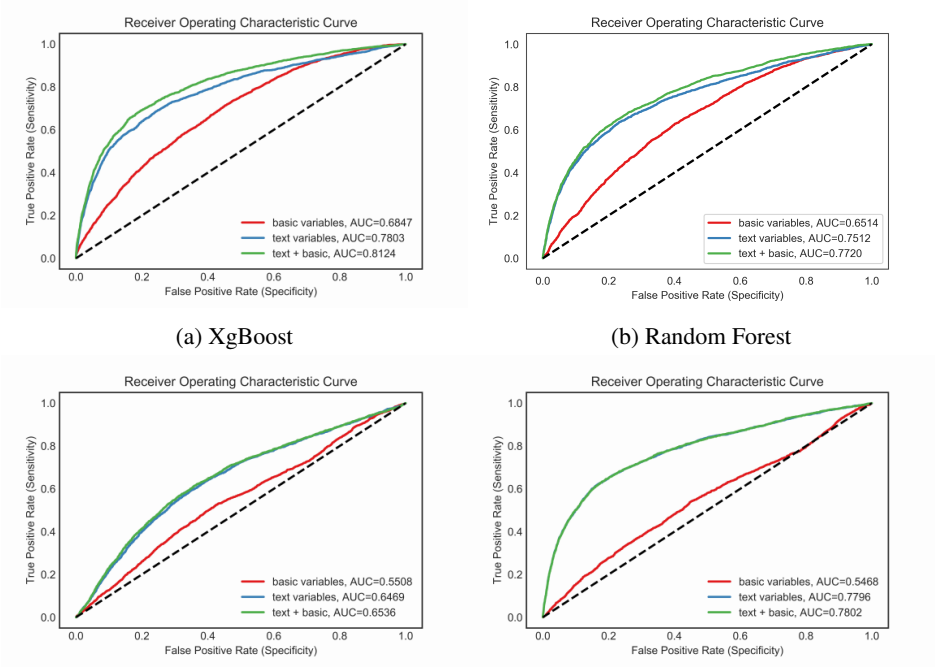
The predicted churn risks are compared against the ground truth labels to compute the AUC scores as showed in Table 8.

The results in Table 8 show that our proposed models are effective for all four datasets with different features. Especially in the case of investment data, the prediction accuracy improves substantially with an average AUC scores increase ranging from 10.28% to 23.34%. Due to the high competition in the financial services industry, "Investment" clients normally have lower levels of attachment or loyalty to the firm and churn for other companies to gain more financial benefits. Table 6 confirms that the customers in the Investment dataset are the most likely to churn. It is thus particularly valuable for the company to retain "Investment" customers as they are generating higher revenue than other types of clients. Thereafter, we mainly focus on the analysis for these customers. Based on the predicted churn risk, we could help the company develop appropriate retention strategies targeting especially at the highest churn risk clients in the "Investment" datasets. The ROC curve plots indicate that using the multi-stacking feature set can further lift the performance of the forecast model compared to the one using only structured data. The results from models with the four different classification algorithms confirmed the effectiveness of our unstructured data mining approach. Figure 7 visualizes the results using the ROC curves of the Investment model for all algorithms. The prediction performance by merely using the structured data is not good, but using the various textual features or the combination of using both structured and textual features could achieve much better prediction performance.

Logistic regression is the best performing algorithms in terms of AUC score improvement. For Investment dataset, our forecast models achieved a significant increase of up to 23.34% when logistic regression is used. The churn prediction models using basic features cannot perform well in this Investment dataset, and the text models can provide significantly more useful information. The results prove that our advanced multi-stacking features methodology is suitable for the customer churn forecast model in general. The model with the XGB algorithm achieved the best AUC

Table 8
AUC results on the models' prediction accuracy

Models	AUC scores				AUC scores			
	Basic	Text	Stacked	Increment	Basic	Text	Stacked	Increment
	Non-employer Super				Employer Super			
XGB	0.7261	0.7229	0.7818	5.56%	0.7980	0.7562	0.8378	3.98%
RF	0.7056	0.7323	0.7572	5.16%	0.7746	0.7465	0.7930	1.84%
LR	0.6024	0.7226	0.7329	13.05%	0.6508	0.7572	0.7699	11.91%
NB	0.6094	0.6451	0.6526	4.32%	0.6756	0.6402	0.6756	0%
	Pension				Investment			
XGB	0.7843	0.7245	0.8220	3.77%	0.6847	0.7803	0.8124	12.77%
RF	0.6932	0.7518	0.7723	7.91%	0.6514	0.7512	0.7720	12.06%
LR	0.6567	0.7143	0.7336	7.69%	0.5468	0.7796	0.7802	23.34%
NB	0.6500	0.6893	0.6985	4.85%	0.5508	0.6469	0.6536	10.28%



(a) XgBoost (b) Random Forest

(c) Gaussian Naive Bayes (d) Logistic Regression

Figure 7: Multi-stacking Ensemble Churn Prediction Model for Investment dataset

score of 0.8124. Therefore, we use the predicted churn risk from this model to perform further customer segmentation and analysis for retention strategies.

We also built and compared models with BERT and BiLSTM: (1) TF-IDF + XgBoost, (2) Word2Vec + XgBoost, (3) BERT Embedding + XgBoost, (4) Word2Vec + BiLSTM, and (5) BERT Embedding + BiLSTM. We compared the results of (6) our final model against those of the multi-stacking ensemble models (7) using BERT as the only text features and (8) adding BERT to our current stacked text. The results in Table 9 for the Investment dataset showed that the models with BERT do not outperform our model. We can conclude that we have found a relatively suitable and optimal solution for our research context.

Table 9
Compare our approach with BERT embedding and BiLSTM for Investment dataset

Model	Algorithm	Multi-Stacking	Features	AUC Score
(1)	XGB	No	TI	0.6878
(2)	XGB	No	PE	0.6370
(3)	XGB	No	BERT	0.6341
(4)	BiLSTM	No	PE	0.6330
(5)	BiLSTM	No	BERT	0.6537
(6)	XGB	Yes	CP + TI, PE, LI, PT	**0.8124**
(7)	XGB	Yes	CP + BERT	0.7420
(8)	XGB	Yes	CP + BERT, TI, PE, LI, PT	0.7980

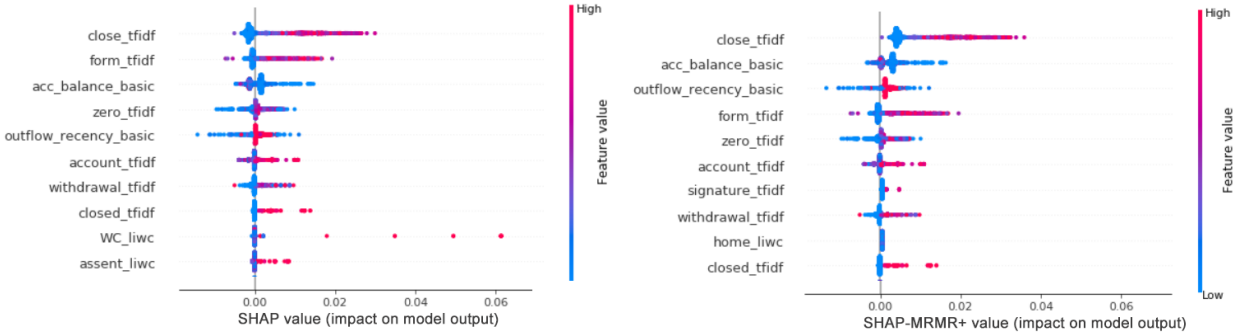


Figure 8: Feature impacts with SHAP and SHAP-MRMR+

4.4. Interpreting Machine Learning Results

We first compare our SHAP-MRMR+ with the original SHAP value for all labeled churn customers to evaluate the feature importance at a global level for the whole dataset. Figure 8 showed the difference in the evaluation of feature impacts between SHAP and SHAP-MRMR+ for the top 10 features. As we can see, the standard SHAP values of "WC_liwc" are scattered in four dots in extremely high values. These are the edge cases where the calls are much longer and the numbers of word count in the transcript are higher. This might introduce bias in interpreting our model. SHAP-MRMR+ lowers the ranking of this kind of features and therefore, it is better than the original SHAP approach in terms of interpreting the prediction model and explaining customer insights. The SHAP-MRMR+ gives more weights to globally important features such as "account balance" and "outflow recency" from the basic customer profile feature set. Hence, these features have a higher impact ranks compared to those in the SHAP values ranking. Our SHAP-MRMR+ gives a lower rank for less meaningful text features such as word count ("WC_liwc") and remove it from the top 10. According to the results of SHAP-MRMR+, seven out of the ten most impactful features are textual features from LIWC and TF-IDF feature set, suggesting that textual information is very useful for the churn prediction model and needs to be taken into account.

In order to visualize SHAP-MRMR+ and interpret our model for each customer segment, we use the force plot [31] where the red color indicates important features which can help to increase the prediction accuracy and the blue color indicates less useful features which might add noise and decrease model value. Figure 9 shows the results on two customer segments: high account balance (top 10%) and low account balance (bottom 10%). Our model suggests that the most impactful features to predict churn for customers with a high account balance are "form" and "credited" from TF-IDF feature set. We can infer from this result that the churn risks of "richer" investors are not dependable on their "outflow recency" and "account balance" (not very impactful features in blue color). However, they might be more concerned with the investment process and probably ask more about "form" and "credited". Meanwhile, the churn risks of investors with a low account balance are still predictable by "account balance" and related text features such as "close" from TF-IDF and "money" from LIWC feature sets. This insightful finding is not easy to identify without using the interpretable machine learning technique like SHAP-MRMR+ which we use in our research.

With the increase of churn prediction accuracy, the company could potentially save around five million dollars in

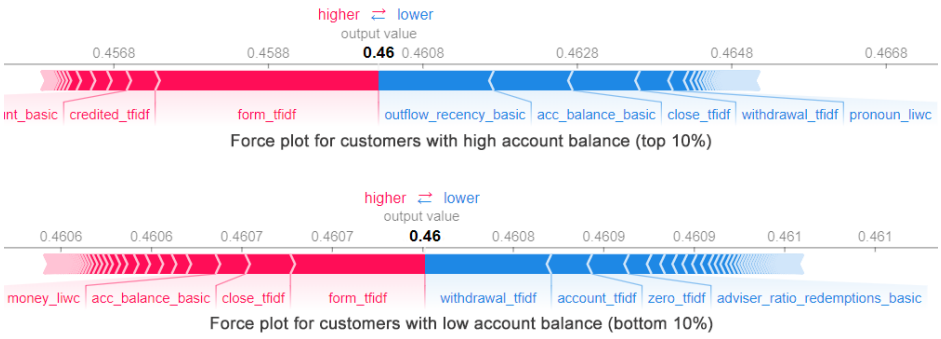


Figure 9: Feature impacts on churn prediction for customer segments with high/low account balance

Figure 10: Personality impacts on churn prediction for individual customer A and customer B

annual revenue just by targeting the top 10% of customers with high churn risks. Therefore, our model incorporates the multi-filtering technique to perform customer segmentation based on the probabilities of their churn decision.

We conduct a descriptive statistical analysis on customer groups with the top 10% highest or lowest scores for each personality trait. The churn risks in Table 10 show that "Investment" customers with lower "Conscientiousness", "Agreeableness" and "Neuroticism" are more likely to leave the company. The finding aligns with previous research on personality traits of the retail customers [7] which shows that "Conscientiousness", "Agreeableness" and "Neuroticism" are significantly correlated with customer empowerment and satisfaction.

Table 10
Predicted churn risk (%) for Customers with Top and Bottom 10% Personality Rank

Segments	Bottom 10%	Top 10%
Openness	7.91	7.22
Conscientiousness	8.94	6.61
Extroversion	7.91	7.34
Agreeableness	9.36	7.67
Neuroticism	8.22	7.16

Moreover, our interpretable machine learning approach can be used to explain the features at the individual customer level. Particularly, the financial services firm is interested in learning the personality traits of each churn customer and leveraging that in their retention strategies. Figure 10 shows the individual force plot for two churn customer A and B. "Conscientiousness", "Agreeableness", and "Neuroticism" are impactful predictors for customer A, whereas "Neuroticism" alone is a strong enough predictor for customer B. These customers with such a special personality are

harder for the financial service firms to retain using the same retention strategies as other customer segments.

4.5. Robustness Analysis

4.5.1. Imbalance learning test

The focal Investment dataset is imbalanced with the number of churn customers are accounted for about 10% in total. In our methodology, we used the raw data as input for the prediction model and defined the class weights accordingly in the machine learning algorithm XgBoost. We test the efficiency of this approach by comparing the prediction results with other commonly used methods in handling imbalance data, e.g. over-sampling and under-sampling methods such as SMOTE, SMOTEENN, SMOTETomek, SVM-SMOTE, and Borderline-SMOTE [29]. The test set ratio for this experiment is 20%, where we use 80% of the data for training using SMOTE methods, and test the prediction models on the out-of-bag 20% of data. The results in Table 11 show that our raw data with class weights achieve higher AUC scores. Since the XgBoost algorithm already assigns different weights to handle the imbalance class, using SMOTE methods to generate synthetic data might inject noise for the machine learning algorithm. Due to the nature of our application, the business focuses the top of the churn customer ranking list, thus AUC is a suitable choice for this study. However, the prediction model with SMOTE methods might achieve better results in other metrics such as F1 score, e.g. the F1 score of the SMOTEENN method is higher than that of the raw data (See Table 11).

Table 11
Evaluation of prediction models using raw and SMOTE transformed data

	AUC	F1 Score
Raw Data	0.7464	0.7421
SMOTE	0.6919	0.7032
SVMSMOTE	0.6840	0.7087
BorderlineSMOTE	0.7006	0.7156
SMOTEENN	0.7074	0.8249
SMOTETomek	0.6951	0.7204

4.5.2. Test of statistical significance

We performed the pair-wise t-test to check the performance of our prediction model using XgBoost compared to the other three algorithms, including Gaussian Naïve Bayes, Logistic Regression, and Random Forest. We conducted the test on the overall "AUC", "accuracy", "precision", "recall" and "F1" scores, and found that our model is statistically better than others at 95% confidence level. Table 12 reports the statistics values and p-values of the pair-wise t-test. Particularly the results of the "AUC" t-test are significant in our final model, which suits the firm's interest in the customer churn probability.

Table 12
Pair-wise t-test on model performance (p-values in brackets)

	Gaussian Naive Bayes	Logistic Regression	Random Forest
AUC	16.3748 (<0.00001)	14.6460 (<0.00001)	7.2400 (0.00003)
Accuracy	102.1689 (<0.00001)	96.5782 (0.00209)	78.2965 (0.00091)
Precision	7.8210 (0.00189)	10.0824 (0.00040)	4.1975 (0.00344)
Recall	15.0128 (<0.00001)	8.1163 (0.00026)	4.8814 (<0.00001)
F1	5.1293 (<0.00001)	5.0790 (0.00025)	4.9702 (<0.00001)

Note: Results are reported as significant at 95% confidence level.

As we focus on using the predicted personality traits as both the input for our churn forecast model and customer segmentation for marketing strategies, we want to test the statistical significance of all five traits. We are using one-way ANOVA [16] and Tukey's HSD [50] tests with the same null hypothesis $H0$: The personality traits are not statistically different. As we can see from Table 13 and Table 14, the test p-values are significant ($p < 0.05$) at 95% confidence level. We can reject $H0$ and conclude that all predicted personality traits are statistically different.

Table 13
One-way ANOVA test on the statistical difference of personality traits

Source	Sum of Squares	Degree of Freedom	F statistics	p-value
C(treatments)	110.4435	4.0	345808.8245	<0.00001
Residual	9.3502	117105.0		

Note: Results are reported as significant at 95% confidence level.

Table 14
Turkey's HSD test on the statistical difference of pairwise personality traits

Trait 1	Trait 2	meandiff	p-value	lower	upper	reject H0
agreeableness	conscientiousness	-0.0159	0.001	-0.0161	-0.0156	True
agreeableness	extraversion	-0.0691	0.001	-0.0693	-0.0689	True
agreeableness	neuroticism	-0.0058	0.001	-0.006	-0.0055	True
agreeableness	openness	0.0236	0.001	0.0234	0.0239	True
conscientiousness	extraversion	-0.0532	0.001	-0.0534	-0.053	True
conscientiousness	neuroticism	0.0101	0.001	0.0099	0.0103	True
conscientiousness	openness	0.0395	0.001	0.0393	0.0397	True
extraversion	neuroticism	0.0633	0.001	0.0631	0.0635	True
extraversion	openness	0.0927	0.001	0.0925	0.0929	True
neuroticism	openness	0.0294	0.001	0.0292	0.0296	True

Note: Results are reported as significant at 95% confidence level.

4.5.3. Hyper-parameters tuning

We evaluated our model under various hyper-parameter settings. The list of tested hyper-parameters included: minimum sum of instance weight needed in a child (min_child_weight), minimum loss reduction required to make a further partition on a leaf node of the tree (gamma), subsample ratio of the training instances (subsample), subsample ratio of columns when constructing each tree (colsample_bytree), maximum depth of a tree (max_depth), and L1 regularization term on weights (reg_alpha). The results in Table 15 show the AUC scores for models with different settings and the final column indicates the best setting which we used in our final prediction model. Overall, the AUC scores did not fluctuate much under different hyper-parameter settings. This suggests that our model is robust.

Table 15
AUC results with different hyper-parameter settings

		AUC Scores			
Hyper-parameter	Tested Settings	1	2	3	Final setting
min_child_weight	[**1**, 5, 10]	**0.7312**	0.7293	0.7257	1
gamma	[**0.5**, 1, 2]	**0.7299**	0.7291	0.7211	0.5
subsample	[0.6, 0.8, **1**]	0.7263	0.7290	**0.7312**	1
colsample_bytree	[**0.6**, 0.8, 1]	**0.7313**	0.7307	0.7312	0.6
max_depth	[3, 4, **5**]	0.7312	0.7361	**0.7362**	5
reg_alpha	[0.001, **0.01**, 0.1]	0.7301	**0.7302**	0.7288	0.01

4.6. Discussion

The empirical results have proven our hypothesis that unstructured data is useful for customer churn forecast. Using text features have significantly improved our prediction accuracy measured by AUC scores, Text data is also insightful to extract customer personalities and characteristics for further analysis. Within the scope of this paper, we only researched the customer call logs data. In the future, different types of text data such as client emails or social media posts might be leveraged to further extend our customer understanding for better services.

The limitation of our research is that due to customer privacy, we do not have direct access to the recorded calls.

The poor quality of text transcription services from the third party has made the feature extraction using advanced embedding models ineffectively. Therefore, we are focusing on more term-based approaches to extract text features.

After customer segmentation, we have obtained profiles and characteristics of customers with high churn risks. Customers, who generally have a higher churn probability, often have a low account balance with negative balance change, which means they often withdraw their funds. The lower churn risk customers also have a higher account tenure, which means they are long-term clients and unlikely to churn. clients with high churn risk also call in more frequently and have lower "Conscientiousness", "Agreeableness" and "Neuroticism".

Based on these findings, we propose appropriate targeted marketing strategies aiming at those customers with high churn risks, especially with their individual personality profiles, to increase the engagement and loyalty of these customers. We analyze some commonly used retention strategies and adapt them to suit our customer personality profiles in order to maximize the effectiveness of each marketing campaign. We separate these strategies into two categories for better analysis: direct and indirect marketing.

Direct Marketing

As the cost of direct marketing is much lower than the indirect one, companies tend to use these strategies more often. Unfortunately, in our case, the "Agreeableness" personality ranks of customers with high churn risks are generally lower, and therefore they may be less likely to respond to direct marketing strategies, including telemarketing and commercial advertising. However, research has shown that customers are loyal to brands that have similar personality traits to themselves [56]. Based on this finding, we can tailor a marketing campaign that is personalized to their personality preferences. The image and text used in any communication channel can be personalized to represent their personality, e.g., these customers would prefer to look at the advertisement with a portrait of someone who looks calm, confident, and self-loving.

On the other hand, these customers are independent thinkers and do not like to take advice. The products and services marketing should present to them as many investment options as possible. For example, the advisor can suggest the customers invest in emerging tech stocks that are risky but have more familiar products with them, rather than some boring government bonds. As they are self-satisfied and self-indulgent, they can also be more responsive to the reward-based loyalty program, e.g., special fee reduction for investments on their birthdays.

Indirect Marketing

Indirect marketing is often costly but more effective for consumers with a low rank of "Conscientiousness", "Agreeableness" and "Neuroticism". Moreover, research has shown that consumers tend to take recommendations from others with the same personalities than with opposite characteristics [2]. These findings help us suggest appropriate strategies for relationship marketing via word-of-mouth. We can suggest the company to identify long-term customers who have similar personality profiles and invite them to become brand ambassadors. With an incentive scheme for motivation, they can help spread the words about the firm to friends and family who have similar personalities. This strategy will be not only beneficial for customer retention but also can potentially attract new investors for the company.

Besides, indirect marketing via social media channels is one of the most cost-effective indirect marketing approaches in recent years. Knowing a customer trait based on text mining from their social media posts will be a huge advantage for customer services. For example, a marketing agent with the same low level of "Neuroticism" will reply to calls from customers with this trait. As people are more comfortable talking to others with similar traits who use similar expressions, their brand engagement and loyal might increase even further in this case.

5. Conclusions

Inside the extent of this paper, we have adopted an unconventional strategy when utilizing unstructured information from client call logs to construct a multi-stacking ensemble churn prediction model and segmenting customers using an interpretable machine learning approach on their profiles and personalities. Overall, the results from the experiments show that the unstructured data, e.g., customer call logs, can be used to generate meaningful insights, and interpretable machine learning should be utilized in all types of customer information systems. The Superannuation firm can potentially save millions of dollars in profit by early identifying high churn risk clients and personalizing marketing strategies to achieve a customer retention rate. The customer analytics field propels giving organizations more intends to comprehend their client on a more extensive quantitative premise utilizing distinctive sort of information as opposed to customary ones.

This is a principal investigation inside the financial services industry in Australia to consolidate both structured and unstructured data to solve problems in customer retention. The empirical experiments demonstrate that unstructured

information contains useful insights that can enhance the precision of the churn risk forecasting on various customer datasets. It does not just help the firm design better-targeted marketing plans for customer retention strategies and save million of dollars in profits yet but also establishes an underlying framework for the utilization of different data types and interpretable machine learning in other information and business intelligence systems. The four datasets with distinguished customer profiles demonstrate our proposed approach would work for different customers, and the method can be generalized for other industries.

References

[1] Adadi, A., Berrada, M., 2018. Peeking inside the black-box: A survey on Explainable Artificial Intelligence (XAI). IEEE Access 6, 52138–52160.

[2] Adamopoulos, P., Ghose, A., Todri, V., 2018. The impact of user personality traits on word of mouth: Text-mining social media platforms. Information Systems Research 29, 612–640.

[3] Al-Hawari, M.A., 2015. How the personality of retail bank customers interferes with the relationship between service quality and loyalty. International Journal of Bank Marketing 33, 41–57.

[4] Ali, Ö.G., Arıtürk, U., 2014. Dynamic churn prediction framework with more effective use of rare event data: The case of private banking. Expert Systems with Applications 41, 7889–7903.

[5] Almana, A.M., Aksoy, M.S., Alzahrani, R., 2014. A survey on data mining techniques in customer churn analysis for telecom industry. International Journal of Engineering Research and Applications 45, 165–171.

[6] Biel, J.I., Gatica-Perez, D., 2013. The youtube lens: Crowdsourced personality impressions and audiovisual analysis of vlogs. Multimedia, IEEE Transactions on 15, 41–55.

[7] Castillo, J., 2017. The relationship between big five personality traits, customer empowerment and customer satisfaction in the retail industry. Journal of Business and Retail Management Research (JBRMR) 11.

[8] Chen, T., Guestrin, C., 2016. Xgboost: A scalable tree boosting system, in: Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining, ACM. pp. 785–794.

[9] Chu, C., Xu, G., Brownlow, J., Fu, B., 2016. Deployment of churn prediction model in financial services industry, in: 2016 International Conference on Behavioral, Economic and Socio-cultural Computing (BESC), IEEE. pp. 1–2.

[10] Costa, P.T., McCrae, R.R., 1992. Four ways five factors are basic. Personality and individual differences 13, 653–665.

[11] Coussement, K., Van den Poel, D., 2008. Integrating the voice of customers through call center emails into a decision support system for churn prediction. Information & Management 45, 164–174.

[12] Coussement, K., Van den Poel, D., 2009. Improving customer attrition prediction by integrating emotions from client/company interaction emails and evaluating multiple classifiers. Expert Systems with Applications 36, 6127–6134.

[13] De Bock, K.W., Van den Poel, D., 2012. Reconciling performance and interpretability in customer churn prediction using ensemble learning based on generalized additive models. Expert Systems with Applications 39, 6816–6826.

[14] Devlin, J., Chang, M.W., Lee, K., Toutanova, K., 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805 .

[15] Farquad, M.A.H., Ravi, V., Raju, S.B., 2014. Churn prediction using comprehensible support vector machine: An analytical CRM application. Applied Soft Computing 19, 31–40.

[16] Fisher, R.A., et al., 1950. Statistical methods for research workers. Statistical methods for research workers. .

[17] Fornell, C., Wernerfelt, B., 1987. Defensive marketing strategy by customer complaint management: a theoretical analysis. Journal of Marketing research 24, 337–346.

[18] Friedman, J.H., 2001. Greedy function approximation: a gradient boosting machine. Annals of statistics , 1189–1232.

[19] Goldberg, L.R., 1990. An alternative" description of personality": the Big-Five factor structure. Journal of personality and social psychology 59, 1216.

[20] Grover, V., Chiang, R.H., Liang, T.P., Zhang, D., 2018. Creating strategic business value from big data analytics: A research framework. Journal of Management Information Systems 35, 388–423.

[21] Huang, Y., Zhu, F., Yuan, M., Deng, K., Li, Y., Ni, B., Dai, W., Yang, Q., Zeng, J., 2015. Telco churn prediction with big data, in: Proceedings of the 2015 ACM SIGMOD international conference on management of data, ACM. pp. 607–618.

[22] Hung, S.Y., Yen, D.C., Wang, H.Y., 2006. Applying data mining to telecom churn management. Expert Systems with Applications 31, 515–524.

[23] Karahoca, A., Bilgen, O., Karahoca, D., 2016. Churn management of e-banking customers by fuzzy AHP, in: Handbook of Research on Financial and Banking Crisis Prediction Through Early Warning Systems. IGI Global, pp. 155–172.

[24] Kendall, M.G., 1938. A new measure of rank correlation. Biometrika 30, 81–93.

[25] Keramati, A., Ghaneei, H., Mirmohammadi, S.M., 2016. Developing a prediction model for customer churn from electronic banking services using data mining. Financial Innovation 2, 10.

[26] Kim, S.H., Kim, M., Holland, S., 2018. How customer personality traits influence brand loyalty in the coffee shop industry: the moderating role of business types. International journal of hospitality & tourism administration 19, 311–335.

[27] Kitchens, B., Dobolyi, D., Li, J., Abbasi, A., 2018. Advanced customer analytics: Strategic value through integration of relationship-oriented big data. Journal Of Management Information Systems 35, 540–574.

[28] Kosinski, M., Matz, S.C., Gosling, S.D., Popov, V., Stillwell, D., 2015. Facebook as a research tool for the social sciences: Opportunities, challenges, ethical considerations, and practical guidelines. American Psychologist 70, 543.

[29] Lemaître, G., Nogueira, F., Aridas, C.K., 2017. Imbalanced-learn: A python toolbox to tackle the curse of imbalanced datasets in machine learning. The Journal of Machine Learning Research 18, 559–563.
[30] Lundberg, S.M., Lee, S.I., 2017. A unified approach to interpreting model predictions, in: Advances in Neural Information Processing Systems, pp. 4765–4774.
[31] Lundberg, S.M., Nair, B., Vavilala, M.S., Horibe, M., Eisses, M.J., Adams, T., Liston, D.E., Low, D.K.W., Newman, S.F., Kim, J., et al., 2018. Explainable machine-learning predictions for the prevention of hypoxaemia during surgery. Nature biomedical engineering 2, 749.
[32] Mairesse, F., Walker, M., 2007. PERSONAGE: Personality generation for dialogue, in: Annual Meeting-Association For Computational Linguistics, p. 496.
[33] Manner, C.K., 2017. Who posts online customer reviews? the role of sociodemographics and personality traits. Journal of Consumer Satisfaction, Dissatisfaction and Complaining Behavior 30, 23–23.
[34] Pearson, K., 1895. Note on regression and inheritance in the case of two parents. Proceedings of the Royal Society of London 58, 240–242.
[35] Peng, H., Long, F., Ding, C., 2005. Feature selection based on mutual information: criteria of max-dependency, max-relevance, and min-redundancy. IEEE Transactions on Pattern Analysis & Machine Intelligence , 1226–1238.
[36] Pennebaker, J.W., Francis, M.E., Booth, R.J., 2001. Linguistic inquiry and word count: Liwc 2001. Mahway: Lawrence Erlbaum Associates 71, 2001.
[37] Pennebaker, J.W., King, L.A., 1999. Linguistic styles: language use as an individual difference. Journal of personality and social psychology 77, 1296.
[38] Ponce-López, V., Chen, B., Oliu, M., Corneanu, C., Clapés, A., Guyon, I., Baró, X., Escalante, H.J., Escalera, S., 2016. Chalearn lap 2016: First round challenge on first impressions-dataset and results, in: European Conference on Computer Vision, Springer. pp. 400–418.
[39] Ravi, K., Ravi, V., Prasad, P.S.R.K., 2017. Fuzzy formal concept analysis based opinion mining for CRM in financial services. Applied Soft Computing 60, 786–807.
[40] Rehurek, R., Sojka, P., 2010. Software framework for topic modelling with large corpora, in: In Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks, Citeseer.
[41] Reimers, N., Gurevych, I., 2020. Making monolingual sentence embeddings multilingual using knowledge distillation. arXiv preprint arXiv:2004.09813 URL: http://arxiv.org/abs/2004.09813.
[42] Rudin, C., 2019. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. Nature Machine Intelligence 1, 206.
[43] Sarkar, C., Bhatia, S., Agarwal, A., Li, J., 2014. Feature analysis for computational personality recognition using youtube personality data set, in: Proceedings of the 2014 ACM Multi Media on Workshop on Computational Personality Recognition, ACM. pp. 11–14.
[44] Scherer, A., Wünderlich, N.V., Von Wangenheim, F., 2015. The Value of Self-Service: Long-Term Effects of Technology-Based Self-Service Usage on Customer Retention. MIS quarterly 39.
[45] Schuster, M., Paliwal, K.K., 1997. Bidirectional recurrent neural networks. IEEE Transactions on Signal Processing 45, 2673–2681.
[46] Seth, G., 2008. Unstructured data and the 80 percent rule. Breakthrough Analysis - Bridgepoints .
[47] Spearman, C., 1904. The proof and measurement of association between two things. The American journal of psychology 15, 72–101.
[48] Sun, S., Luo, C., Chen, J., 2017. A review of natural language processing techniques for opinion mining systems. Information fusion 36, 10–25.
[49] Sundarkumar, G.G., Ravi, V., 2015. A novel hybrid undersampling method for mining unbalanced datasets in banking and insurance. Engineering Applications of Artificial Intelligence 37, 368–377.
[50] Tukey, J.W., 1949. Comparing individual means in the analysis of variance. Biometrics , 99–114.
[51] Verbeke, W., Martens, D., Mues, C., Baesens, B., 2011. Building comprehensible customer churn prediction models with advanced rule induction techniques. Expert systems with applications 38, 2354–2364.
[52] Vinciarelli, A., Mohammadi, G., 2014. A survey of personality computing. IEEE Transactions on Affective Computing 5, 273–291.
[53] Vo, N.N., Liu, S., Brownlow, J., Chu, C., Culbert, B., Xu, G., 2018a. Client Churn Prediction with Call Log Analysis, in: International Conference on Database Systems for Advanced Applications, Springer. pp. 752–763.
[54] Vo, N.N., Liu, S., He, X., Xu, G., 2018b. Multimodal mixture density boosting network for personality mining, in: Pacific-Asia conference on knowledge discovery and data mining, Springer. pp. 644–655.
[55] Wei, C.P., Chiu, I.T., 2002. Turning telecommunications call details to churn prediction: a data mining approach. Expert systems with applications 23, 103–112.
[56] Yao, Q., Chen, R., Xu, X., 2015. Consistency between consumer personality and brand personality influences brand attachment. Social Behavior and Personality: an international journal 43, 1419–1427.
[57] Yee Liau, B., Pei Tan, P., 2014. Gaining customer knowledge in low cost airlines through text mining. Industrial management & data systems 114, 1344–1359.
[58] Zhou, S., Qiao, Z., Du, Q., Wang, G.A., Fan, W., Yan, X., 2018. Measuring customer agility from online reviews using big data text analytics. Journal of Management Information Systems 35, 510–539.