GPT-LGBM: A ChatGPT-based integrated framework for credit scoring with textual and structured data

Li Yu[a], Xuefei Bai[a], Zhiwei Chen[a, *]

[a] Shanghai Key Laboratory of Financial Information Technology, Shanghai University of Finance and Economics, Shanghai 200433, People's Republic of China

[*] Corresponding author

Li Yu
yu.li@mail.shufe.edu.cn
Shanghai Key Laboratory of Financial Information Technology, Shanghai University of Finance and Economics, Shanghai 200433, People's Republic of China

Xuefei Bai
bai.xuefei@163.sufe.edu.cn
Shanghai Key Laboratory of Financial Information Technology, Shanghai University of Finance and Economics, Shanghai 200433, People's Republic of China

Zhiwei Chen (corresponding author)
ycczw2018@163.sufe.edu.cn
Shanghai Key Laboratory of Financial Information Technology, Shanghai University of Finance and Economics, Shanghai 200433, People's Republic of China

GPT-LGBM: A ChatGPT-based integrated framework for credit scoring with textual and structured data

Abstract

With the rapid growth of the credit market, credit scoring becomes increasingly important for credit risk management. Current credit scoring models tend to use both structured data and textual data for credit assessment, and the main difficulty lies in extracting meaningful textual features from unstructured loan texts. ChatGPT is a generative large language model that can proficiently comprehend textual material and perform reasoning tasks. Therefore, this study proposes a ChatGPT-based integrated framework, named as GPT-LGBM. The proposed framework can be divided into two stages: the initial stage involves psychological feature extraction with ChatGPT, followed by credit risk classification using light gradient boosting machine (LightGBM). During the first stage, two paths are designed to extract borrowers' Big Five personality traits from loan texts without manual annotation. One denoted as Path-D employs ChatGPT directly to analyze all loan texts, which aims at maximizing the benefits of using ChatGPT. The other denoted as Path-K adopts the idea of knowledge distillation to combine ChatGPT and the deep learning model RoBERTa, which can mitigate the risk of data breach by restricting ChatGPT's access to all loan texts. In the second stage, the Big Five personality traits and structured features are concatenated as inputs for the LightGBM model. Extensive experiments based on the Lending Club data demonstrate that GPT-LGBM is an effective framework for credit evaluation.

Keywords: Credit scoring, ChatGPT, Knowledge distillation, Big Five personality traits, LightGBM

1. Introduction

Credit scoring is an economic activity that classifies a loan applicant into different classes according to their likely repayment behavior, e.g., non-default ones that are expected to repay on time and default ones that are expected to fail [1,2]. Since small improvements in prediction of borrowers' credibility can provide great gains in economic benefits, lenders make every effort to identify potential defaulters [3–5]. However, it is quite a challenging task to evaluate the possibility of defaulting due to the severe information asymmetry between borrowers and lenders [6]. As commonly observed, borrowers are more aware of their credibility than lenders [7]. To mitigate information asymmetry risks, lenders will typically construct credit scoring models to establish the mapping relationship between available borrower data and their own credit quality.

While previous researches on credit scoring have focused on structured data, recent researches shed light on the significant role that unstructured textual data can play in loan default prediction. Here, structured data includes borrower characteristics, financial indicators, macroeconomic indicators, etc., while unstructured textual data refers to textual loan descriptions, social media postings, etc. [6–9]. Several studies support the notion that unstructured textual data can offer insights into the borrowers' creditworthiness from a different angle, and combining the two information profiles can improve the performance of credit scoring models [10–12].

There are three major approaches to retrieve textual features from unstructured texts: rule-based statistical methods, clustering-based machine learning methods and deep learning methods. Based on manually defined statistical indicators, rule-based statistical methods can easily extract textual features, such as word frequency and the number of punctuation marks [12,13]. But these features often cannot accurately convey the semantics of the text and the quality of the designed rules will largely determine the efficacy of the methods. Clustering-based machine learning methods, such as Latent Dirichlet allocation topic (LDA), are good at clustering similar words and finding hidden topics, but the results are sensitive to the number of topics and may have semantic overlap [10,14]. Deep learning methods demonstrate satisfactory performance, yet they suffer from limited interpretability. To address this problem, psychological features, such as Big Five personality traits, are introduced for credit risk evaluation [7,9]. These psychological characteristics improve both the predictive performance and interpretability of the credit scoring model, but it requires considerable manual effort to obtaining psychological features through questionnaire surveys, expert labeling, etc.

ChatGPT, released by OpenAI, is an artificial intelligence language model that can generate convincing responses to users on various topics [15,16]. By autonomously learning from massive data, ChatGPT can be applied to various domains such as education, healthcare, finance, e-commerce, etc. [17]. Many articles demonstrate the superiority of ChatGPT in reasoning and annotating tasks [18,19], encompassing the field of Big Five personality traits prediction [20,21]. Hence, the potential exists for ChatGPT to increase the prediction performance of credit scoring models with both structured and textual data.

Based on the review of the available literature, ChatGPT has not been included in the existing credit scoring models. It is necessary to explore the most effective means of implementing ChatGPT in the credit assessment field. However, data security is a salient concern for financial institutions to minimize the disclosure of borrower data [22,23]. Recent news of Samsung Electronics Co. reveals that utilizing generative AI platforms is at risk of data leakage [24]. Although ChatGPT has made every effort to minimize data exposure, there may still be potential risks. Therefore, it is also important to strike a balance between the benefits of utilizing ChatGPT and the risk of data exposure.

To fill these research gaps, the proposed research employs ChatGPT and Light Gradient Boosting Machine (LightGBM) for constructing an integrated framework for credit scoring, which is named as GPT-LGBM. The whole framework can be divided into two consecutive stages. In the first stage, based on the designed prompts, ChatGPT is applied to transform credit-related textual data into 5 psychological features, i.e., Big Five personality traits. In this stage, two alternate paths named Path-D and Path-K are developed according to different data protection considerations. Path-D directly utilizes ChatGPT to convert all textual data into Big Five personality traits, which can maximize the benefits of using ChatGPT. Path-K follows the idea of knowledge distillation for textual data transformation. Along this path, ChatGPT acts as the teacher model and processes partial textual data, whereas the deep learning model RoBERTa acts as the student model and learns from the ChatGPT-generated responses. In the second stage, by combining the obtained Big Five personality traits and the structured features, concatenated vectors are generated and fed into LightGBM for credit risk classification. To illustrate the effectiveness of the proposed framework, extensive experiments are conducted on an open-access dataset from a well-known P2P platform Lending Club. The results regarding four performance metrics show that the proposed GPT-LGBM is a promising framework for credit scoring.

The key contributions of the proposed work can be summarized as two aspects. (1) This paper develops a two-stage framework with ChatGPT and LightGBM. To our knowledge, it is the first study that focuses on the integrated implementation of ChatGPT in the area of credit scoring. ChatGPT is utilized to analyze textual data from loan applicants and infer their personality traits, which can greatly reduce the cost of manual annotation. Since the textual data is transformed into psychological features rather than uninterpretable features, the interpretability of the subsequent classification model is improved. Then, LightGBM is chosen for defaulter classification due to its high-performance. (2) In our framework, two paths are designed for different data protection requirements. Path-D is quite straightforward and yields promising prediction results, as long as ChatGPT can access all textual data. By restricting ChatGPT to access only partial textual data, Path-K is suitable for the application scenarios that place a high priority on data security. With the assistance of RoBERTa that can be deployed locally by credit risk managers, this path can mitigate the risk of data breach while retaining the accuracy of loan default prediction.

The remainder of the paper is organized as follows. Section 2 reviews the research works that are related to this study. Section 3 describes the framework of GPT-LGBM. Case study and experimental results are given in section 4. The last section contains our conclusions and possible directions of future research.

2. Literature review

Credit scoring is playing an increasingly critical role in modern financial systems by assessing loan applicants' creditworthiness. Initially, scholars focus on analyzing the structured data of applicants using statistical and machine learning methods. Lee et al. used classification and regression tree (CART) and multivariate adaptive regression splines (MARS) for credit scoring [25]. Both methods can be easily implemented and interpreted by credit risk managers. Yeh et al. stated that artificial neural network (ANN) can achieve better predictive accuracy in discriminating credit card clients than other five classification methods, i.e., discriminant analysis (DA), logistic regression (LR), Naïve Bayesian classifier (NB), K-nearest neighbor classifiers (KNN), and classification trees (CTs) [26]. Bhattacharyya et al. evaluated support vector machines (SVM), random forests (RF) and LR for credit card fraud prediction [27]. Based on the experiments with real-life data from an international credit card operation, RF performed best across performance measures.

As ensemble methods demonstrate strong classification capabilities, many researchers tend to use them for credit risk analysis. Lessmann et al. carried out a comparative study of 41 classification models on eight retail credit scoring datasets [28]. The experimental results verified the promising performance of the ensemble classifiers. Xia et al. suggested a sequential ensemble credit scoring scheme with extreme gradient boosting (XGBoost) [29]. And interpretability of XGBoost can be addressed by the feature importance scores. Chang et al. utilized XGBoost to construct a credit risk assessment model for financial institutions [30]. He et al., Xia et al. employed XGBoost as the base classifier to construct the ensemble model [31,32]. Liu et al. adopted LightGBM for credit scoring and incorporated a cost-sensitive loss function to accommodate the misclassification costs and classification hardness [33]. To interpret the prediction results, they applied feature importance and partial dependence plots to seek important credit features.

With the rapid development of Internet Finance and Natural Language Processing (NLP) techniques, scholars focus on improving the performance of classification by incorporating textual data, which can be derived from borrowers' social media posts, electronic loan application forms, etc. In general, existing methods for processing credit-related textual data can be categorized into three main groups: rule-based statistical methods, clustering-based machine learning methods and deep learning methods.

Rule-based statistical methods typically extract textual features with manually defined rules. Dorfleitner et al. employed text-related factors such as text length and spelling errors for default probability prediction in peer-to-peer lending [34]. The experimental results demonstrated that predicting the default probability using rule-based factors was difficult. However, Li et al. indicated that text length is a useful feature for identifying borrowers' creditworthiness [13]. Nguyen and Huynh utilized an emotion dictionary to recognize sentiment polarity in corporate financial reports and introduced sentiment indicators for corporate credit risk prediction [35]. Although rule-based statistical methods are simple and intuitive, they lack an understanding of the meaning of whole sentences and their performance is mostly determined by the quality of the rules.

Clustering-based machine learning methods focus on the extraction of textual features through clustering-like techniques. Guo et al. used an LDA model to extract users' topic distributions from the user-generated social media data for credit scoring [8]. Wang and Xu employed an LDA method to detect automobile insurance fraud, which enabled the extraction of textual features from both fraudulent claims and non-fraudulent claims [10]. The experimental results indicated that the text information was essential and contributed to the increased accuracy of fraud detection. Jiang et al. employed an LDA model to extract six credit-related topics from descriptive loan texts [14]. Liu et al. clustered the loan description information on online lending platforms using an LDA method to identify their loan purposes [36]. Xia et al. employed soft information extracted from narrative data by the K-means-based keyword clustering algorithm to predict loan default [11]. The experimental results suggested the utilization of soft information improved the predictability compared to solely using hard information. Wang et al. employed the GloVe algorithm for mapping the terms in descriptive loan texts to an embedding space [5]. Then spatially adjacent terms were automatically clustered into same semantic cliques, which are used as semantic soft factors. Clustering-based machine learning methods excel at grouping similar lexical items and seeking out the hidden topics without supervision, but also lack the capability of understanding whole sentences. Moreover, their performance is often sensitive towards the number of topics or clusters.

Deep learning methods concentrate on using deep learning models for textual feature extraction. Mai et al. investigated the effectiveness of average embedding model and convolutional neural network (CNN) in analyzing textual disclosures of public firms [37]. Compared to LR, SVM, and RF that utilize TF-IDF features, these two deep learning models excelled in analyzing textual data and achieving more accurate bankruptcy classification. Matin et al. integrated CNN, long short-term memory (LSTM) and attention mechanism, to realize the transformation of the firms' annual reports to representative vectors [38]. Zhang et al. showed through experiments that compared to LDA, CNN and Embedding average (EA), Transformer encoder (TE) is a more effective method for extracting information from loan descriptions [6].

In recent years, pre-trained language models have emerged as the dominant methodology in the domain of credit evaluation to extract textual features. Stevenson et al. employed Bidirectional Encoder Representations from Transformers (BERT) to convert textual loan assessments from loan officers into fixed-length vectors [39]. Huang et al. employed word2vec and BERT to extract textual sentiment indicators from the annual reports of Chinese listed companies [40]. Based on user-generated text from Lending Club, Kriebel and Stitz used default states as target labels and trained six supervised deep learning models (CNN, RNN, CRNN, EA, BERT, and RoBERTa) [12]. The experimental results supported the conclusion that deep learning was an effective technique and textual data was useful for credit risk prediction. While these approaches have demonstrated good classification performance, they offer limited interpretability, and the dimensionality of textual features is often an important hyperparameter to be optimized.

To address these limitations, researches introduced psychological features for credit risk evaluation. Lu et al. proposed a mixed methodology to explain borrowers' default risk using general strain theory based on psychological features [9]. Within the methodology, psychological features are obtained utilizing the questionnaire approach, and six different algorithms such as LSTM, Multi-Layer Perceptron (MLP) were trained for psychological features scoring. In this study psychological features exhibited strong predictive ability and interpretability for credit assessment. Yang et al. proposed an online risk assessment framework PsyCredit, which was composed of deep neural network and a personality mining model [7]. Based on the Big Five personality labels obtained through questionnaires, PsyCredit fine-tuned a pre-trained language model BERT. Then the Big Five personality traits generated by the fine-tuned BERT and the traditional features were fed into a deep neural network to get credit scores. The experimental results verified that leveraging Big Five personality labels of borrowers can improve the performance of credit risk assessment. These methods with psychological features generally yield superior predictive performance and interpretability, albeit at the cost of additional labeling effort, which can be time-consuming and labor-intensive.

3. Research framework

During the loan application period, individual investors and investment institutions make a decision on whether to grant credit to a borrower. The decision-making process is usually modeled as a binary classification problem, where the goal is to identify potential defaulters from all borrowers [6,33,36]. The borrowers likely to default will be rejected to granted a loan to avoid significant losses.

This paper proposes an integrated framework named GPT-LGBM, to accurately predict borrowers' credibility. ChatGPT is integrated with machine learning algorithms for efficiently mining hidden patterns from all credit-related data, including structured and textual data. Due to the challenge of handling structured and textual data simultaneously, it is a common practice to transform textual data into structured data and then integrate it with original structured data for subsequent classification. As shown in Fig. 1, the proposed framework is comprised of two stages, where Stage-1 transforms textual data into psychological features based on ChatGPT, and Stage-2 utilizes LightGBM to classify defaulters and non-defaulters based on the concatenated vectors of psychological features and structured features. During Stage-1, a decision maker can follow two alternative paths for implementing ChatGPT. The first path (Path-D) is processing all textual data with ChatGPT directly, which can achieve promising classification results. Given the need to reduce data exposure, the other path (Path-K) can be considered as a data security protection scheme that restricts ChatGPT's access to all textual data. Along this path, a ChatGPT-based knowledge distillation model is proposed which only needs partial textual data to achieve good predictive results. Here, partial textual data refers to a subset of historical textual data, excluding new borrowers' textual data.

Consequently, the proposed framework involves five major steps for training:

(1) Divide the collected credit data into two groups, structured data and textual data. The structured data are represented as N-dimensional vectors, where N is the number of structured features. Since ChatGPT can accommodate diverse textual content, NLP preprocessing techniques for textual data, such as tokenization, lemmatization, and stop word removal, are no longer necessary.

(2) This step requires determining whether ChatGPT is granted to access all textual data or not. If the answer is Yes, proceed to the step 3. Otherwise, skip to step 4.

(3) Path-D: Sequentially load all textual data into ChatGPT with a zero-shot prompting approach. ChatGPT can automatically summarize Big Five personality traits based on each borrower's textual data, which is discussed in detail in Section 3.1. The ChatGPT-generated responses are then transformed into 5 psychological features, i.e. Big Five personality traits, and skip to step 5.

(4) Path-K: Specify the portion of data that can be input into ChatGPT. Follow the same prompt engineering approach as in step 3, ChatGPT can generate corresponding Big Five personality traits based on this portion of data. Then, the ChatGPT-generated responses are utilized to train a student model in the knowledge distillation model. The knowledge distillation model is described in detail in Section 3.2. After training, the student model can transform the remaining textual data into 5 psychological features.

(5) Concatenate the 5-dimensional psychological vector and N-dimensional structured vector of the same borrower. The concatenated vectors (dimension=N+5) and the borrowers' loan status are used to train a classification algorithm, i.e. LightGBM, which is described in detail in Section 3.3.

After training, the classification algorithm can make accurate predictions of risky loans. When a new applicant applies for a loan, the textual data can be transformed into 5 psychological features by following either Path-D or Path-K. The trained LightGBM model is then used to predict the applicant's risk of default with the concatenated vector.

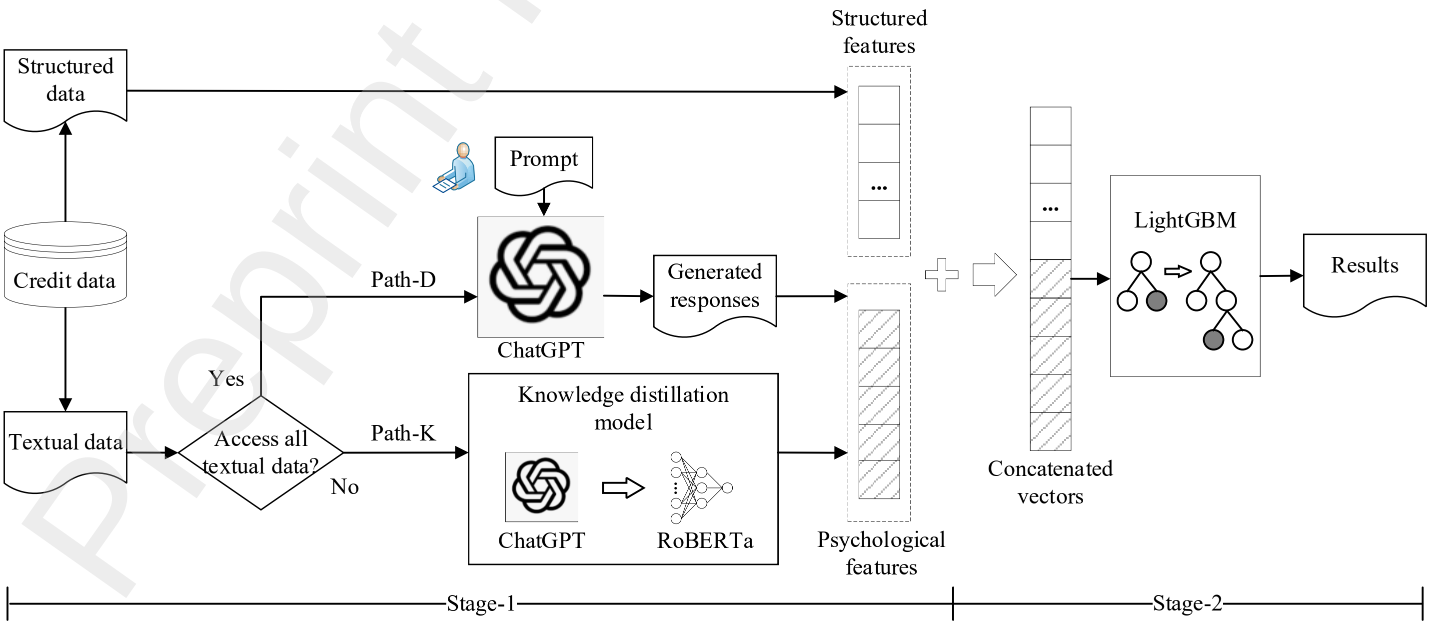

Fig. 1. The framework of GPT-LGBM

3.1 ChatGPT for Big Five personality traits extraction

In the field of credit assessment, research findings indicate that a borrower's credit is affected by his or her psychological characteristics, and the Big Five model is a prevalent and effective method for evaluating individuals' psychological traits [7]. The Big Five model describes five fundamental dimensions of personality traits, namely, openness, conscientiousness, extraversion, agreeableness and neuroticism. As a large-scale generative language model, ChatGPT provides opportunities to accomplish a wide range of natural language processing tasks quickly [15,19]. Therefore, in this study, ChatGPT is utilized to identify personality traits of loan applicants according to their textual data.

ChatGPT accepts a set of instructions, i.e., prompts, and generates responses [41]. Since the performance of ChatGPT is largely affected by prompt engineering [42], it is important to build appropriate prompts based on textual data of loan applicants.

The designed prompt is shown in Fig. 2, which is composed of five major components: Role Injection, Format Description, Task Description, Results Presentation and Borrowers' Text.

(1) The Role Injection component can instruct ChatGPT to act as a specific role type and generate professional responses [43]. To conduct an effective measure of applicants' personality traits, this component gives ChatGPT the role of an expert well versed in both personal credit assessment and psychological assessment.

(2) The Format Description component is a general format requirement. The JSON format is chosen as it is a widely used and user-friendly format.

(3) The Task Description component gives the details of the Big Five personality traits recognition task. Since ChatGPT may conflate neutrality with non-mention, it is required to assess whether a particular personality trait is evident before evaluating its tendency. There are 6 levels for each Big-five personality trait and level 0 stands for non-mention. The last sentence instructs ChatGPT to provide reasons before making any evaluations, which can give ChatGPT time to "think" step by step before drawing a conclusion [44].

(4) The Results Presentation component is a detailed description for representing the results of evaluation. These explicit format guidelines can prevent the generation of irrelevant responses by ChatGPT.

(5) The Borrowers' Text component contains the textual data from each loan applicant. To protect the confidentiality of data, some exact numbers are replaced with asterisks.

When a well-designed prompt is fed into ChatGPT, it can generate the Big Five personality traits of the loan applicant. The ChatGPT-generated response contains levels and reasons regarding Big Five personality traits.

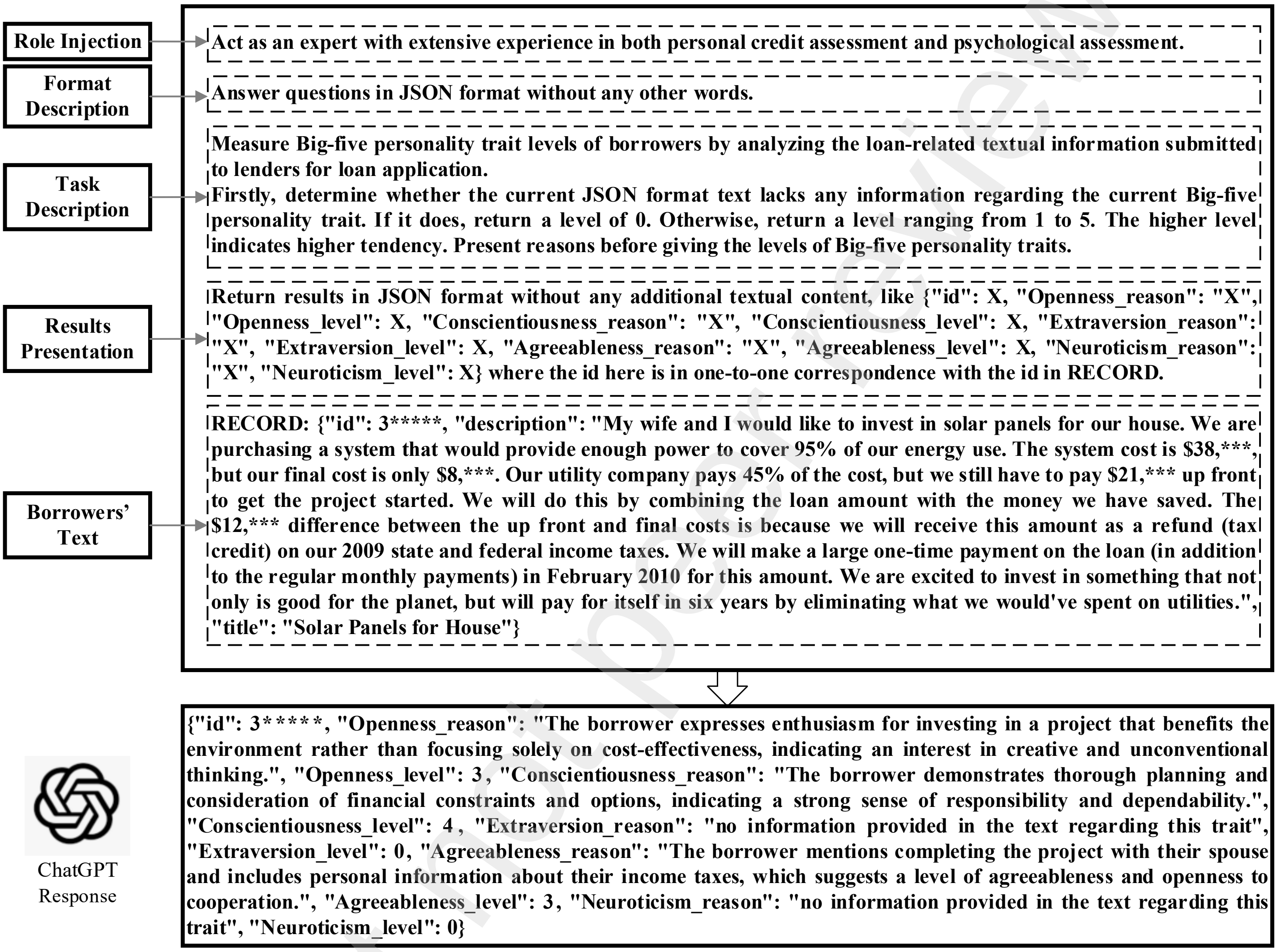

Fig. 2. The structure of a prompt and its corresponding answer

3.2 Knowledge distillation model

In order to protect the privacy and security of data, this paper uses a knowledge distillation model to compress the ChatGPT's Big Five scoring knowledge into a small-scale model [45]. Following the teacher-student network approach for knowledge distillation in Thoker & Gall [46], a straightforward knowledge distillation model based on ChatGPT's knowledge is constructed.

The RoBERTa model is an open-source model that performs well in multiple natural language processing tasks [47]. Compared with artificial general intelligence models such as ChatGPT, it has fewer parameters and is more suitable for deployment by financial institutions. Hence, as shown in Fig. 3, a RoBERTa-based model is proposed as the student model, which maps the loan texts to corresponding Big Five personality traits. The student model is deployed locally to ensure data security. ChatGPT serves as the teacher model. Through knowledge distillation, the knowledge about Big Five scoring in ChatGPT is transferred to the trained student model.

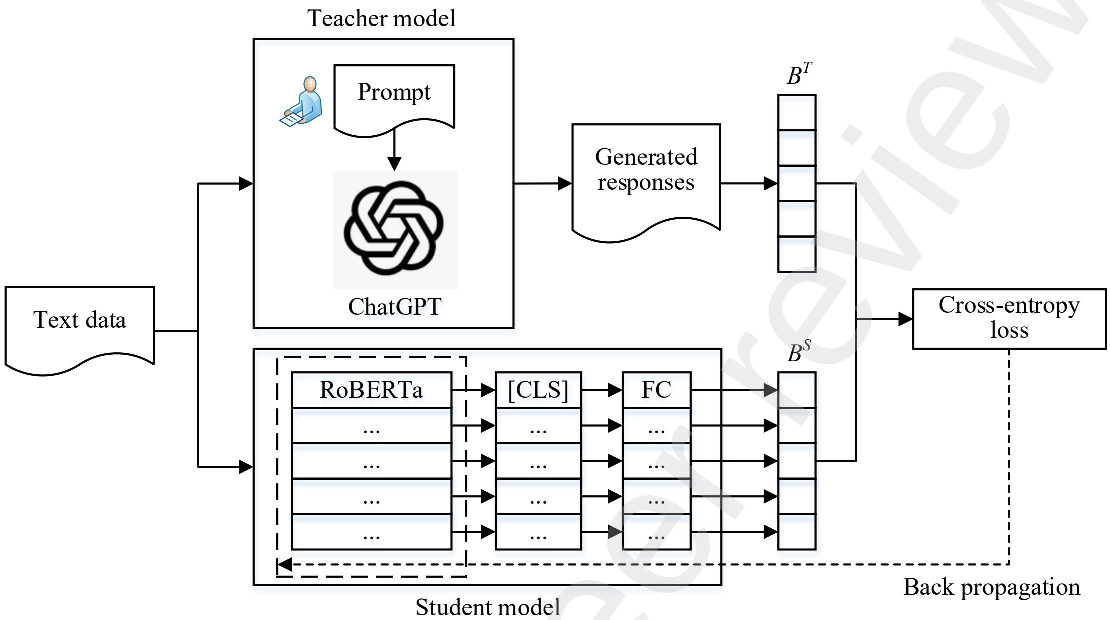

Fig. 3. The framework of the knowledge distillation model

Suppose there are M loan texts $\{t_i\}_{i=1}^{M}$ that ChatGPT can access. Each text sample t_i is embedded in the designed prompt and fed into ChatGPT. Then the Big Five personality traits $\boldsymbol{B}^T = \{\boldsymbol{B}_i^T\}_{i=1}^{M}$ of the loan applicants can be obtained from the generated responses of the teacher model ChatGPT, where $\boldsymbol{B}_i^T = (ope_i^T, con_i^T, ext_i^T, agr_i^T, neu_i^T) \in \mathbb{R}^5$, ope_i^T is the openness level of the ith loan applicant, con_i^T is the conscientiousness level, ext_i^T is the extraversion level, agr_i^T is the agreeableness level and neu_i^T is the neuroticism level. Combining the loan applicants' textual data with the Big Five personality traits generated by the teacher model, a new data set $\boldsymbol{TD} = \{(t_i, \boldsymbol{B}_i^T)\}_{i=1}^{M}$ is built and used for training the student model.

The student model establishes 5 sub-models with the same structure to predict the Big Five personality traits. That is, each personality trait is associated with a sub-model, which consists of a RoBERTa model and a fully connected layer (FC) with 6 output neurons. Every output neuron corresponds to a level of the personality trait. Every RoBERTa model is initialized from a pre-trained RoBERTa model, which employs 12 transformer encoder layers and a hidden size of 768 [47]. The fully connected layer using ReLU activation function is added on the top of the

RoBERTa and the final hidden-state vector of the classification token [CLS] in RoBERTa is fed into this fully connected layer. The output of the fully connected layer represents the probability of the levels for a personality trait. The level with the highest probability is chosen as the predicted level of the personality trait for the ith loan applicant. Through these procedures, the prediction results of the 5 sub-models constitute the Big Five personality traits $\boldsymbol{B}^S = \{\boldsymbol{B}_i^S\}_{i=1}^M$ of the applicants, where $\boldsymbol{B}_i^S = (ope_i^S, con_i^S, ext_i^S, agr_i^S, neu_i^S) \in \mathbb{R}^5$ is the prediction levels of the ith loan applicant from the student model.

Based on cross-entropy loss function, the student model is trained with dataset $\boldsymbol{TD}$ such that $\boldsymbol{B}^S$ can match $\boldsymbol{B}^T$, which enables the student model to acquire Big Five scoring knowledge from the ChatGPT-generated responses. When a borrower submits a new application with textual loan descriptions, the student model is capable of predicting the Big Five levels of the borrower.

3.3 LightGBM for default prediction

As an implementation of the Gradient Boosting Decision Tree (GBDT) algorithm, LightGBM is a state-of-the-art classifier for credit scoring [33]. Compared to black-box classifiers such as SVM and MLP, LightGBM, even as an ensemble model, still preserves the interpretability of tree-based models, which is crucial for credit evaluation.

LightGBM employs both Exclusive Feature Bundling (EFB) and a histogram-based algorithm to process the features. EFB combines sparse variables that are mutually exclusive [48], thereby enhancing the LightGBM's capacity to handle sparse variables when facing large-scale datasets. The histogram-based algorithm enables efficient handling of both continuous and categorical attributes by discretizing them into bins. Differing from other tree models that employ the level-wise tree growing strategies, LightGBM adopts a leaf-wise tree growth approach that circumvents splitting of lower-gain leaf nodes.

As an additive tree-based ensemble model, LightGBM utilizes a forward stagewise algorithm for optimizing the loss function L:

$$L(y, F_t(x)) \approx \sum_{i=1}^{N} \left[L\left(y_i, F_{t-1}(x_i) + g_i f_t(x_i) + \frac{1}{2} h_i f_t^2(x_i)\right)\right] + \Omega(f_t) \qquad (1)$$

where x_i is the concatenated vectors of the ith loan applicant, $y_i \in \{0, 1\}$ represents the status of the ith loan applicant, N is the number of samples, $f_t(x_i)$ represents the predicted value of tth tree for x_i, $F_{t-1}(x_i)$ is the cumulative prediction of the previous $(t-1)$ trees, g_i represents the first derivative of the loss function and h_i represents the second derivative of the loss function. The regularization term $\Omega(f_t)$ is determined by the number of leaf nodes and the values assigned to the leaf nodes [33].

Furthermore, LightGBM uses Gradient-based One-Side Sampling (GOSS) to increase the efficiency of the training process. By sampling instances with varying gradients, GOSS enhances the contribution of samples that have a significant impact on minimizing the loss function value.

Based on the concatenated vectors and loan statuses, LightGBM is trained with the loan application dataset $\boldsymbol{D} = \{(x_{\mathrm{i}},y_{\mathrm{i}})\}_{\mathrm{i}=1}^{N}$. When the psychological features and structured features of a new loan application are inputted into the trained LightGBM model, it can generate a predicted value of default probability $F_T(x)$ based on the ensemble of T decision trees.

4. Experiments and results

4.1 Data

The proposed GPT-LGBM method is evaluated with a large dataset from a famous peer-to-peer (P2P) lending platform, Lending Club. To facilitate borrower-lender matching, Lending Club collects structured and textual data from loan applicants. It also provides free public access to the application and subsequent payment data. As Lending Club data is available since 2007 and its loan descriptions are not available after 2014, the loan records during the period of 2007 to 2014 are collected. Among these records, only the closed loans that reflect complete credit relationships are used. That is, both fully paid and charged-off loans are utilized for empirical analysis.

Similar to Fitzpatrick & Mues [49], the structured data consists of 21 variables, with 14 are from loan listings and the remaining 7 are constructed variables. The 14 variables include annual income (annual_inc) , delinquencies 2 years (delinq_2yrs), debt-to-income ratio (dti), FICO score (fico_range_high), home ownership (home_ownership), inquiries within 6 months (inq_last_6mths), loan amount (loan_amnt), loan term (loan_term), open accounts (open_acc), number of derogatory public records (pub_rec), loan purpose (purpose), revolving balance (revol_bal), revolving line utilization rate (revol_util) and sub grade (sub_grade). The 7 constructed variables are: employment length unknown (emp_length_ukn), the length of employment is known or not (emp_title_ukn), months of credit history (hist_len) , the unemployment rate of USA and the year-on-year change in the OFHEO house price index of the state (hpi_yoychg) , the annual income is verified or not (inc_verified), ratio of installment to total income (inst_to_inc) and the employment title is missing or not (unemp_rate). All categorical variables in string format are transformed into numerical format to meet the input requirements of the machine learning algorithms. The annual income is winsorized at 1% and 99% in order to prevent extreme values from affecting the algorithms.

In each loan record, there are two textual segments, a loan title and a loan description. According to [12], each loan title and its corresponding loan description are combined as one loan text. In addition, automatically created logs are cleaned, and the loan texts with less than 150 tokens are removed. The resulted dataset contains 7239 loan records, of which 84.87% are fully paid and 15.13% are charged-off. Based on statistical analysis, the textual data has an average of 245 tokens, with a maximum of 1078 tokens and a minimum of 150 tokens. 95% of the textual data contains less than 737 tokens.

4.2 Experimental setup

In order to evaluate the ability to distinguish bad loans (charged-off loans) from good loans (fully paid loans), 4 metrics are chosen, i.e. AUC, KS, F1 Score, G-mean. AUC and KS are threshold-independent metrics. F1 Score and G-mean are threshold-dependent metrics and the optimal threshold is obtained through computation of KS in this study. All metrics are calculated based on the confusion matrix. AUC (Area Under ROC Curve) is the area under the receiver operator characteristic curve (ROC curve), which illustrates the performance of an algorithm with various discrimination threshold [23]. The Kolmogorov-Smirnov (KS) statistic is the maximum difference between the cumulative distribution of good loans and bad loans [23]. F1 Score is the harmonic mean of precision and recall. G-mean is the geometric mean of true positive rate and true negative rate.

To fully exploit the dataset and mitigate the impact of randomness, a 10-fold cross-validation process is utilized [7]. All data is stratified split according to the loan status. Moreover, the cross-validation process is repeated three times with three random seeds, and each reported experimental result is the average of 30 times running. Various popular machine learning methods are applied in the experiments as comparative methods, such as LR, KNN, SVM, MLP, DT, RF, GBDT, XGBoost and LightGBM. Bayesian Optimization (BO) is utilized for hyperparameter searching since it can find optimal hyperparameters in a small number of trials [50]. The number of trials is 200 for every algorithm to ensure fair comparisons. The hyperparameter search spaces can be seen in Appendix A.

All the experiments are performed using Python 3.10.11 on a PC with an Intel(R) Core i7–9700K CPU and 64GB RAM, running Ubuntu 22.04 LTS. In order to ensure reproducibility of the results, the versions of Python packages are shown as following: sklearn (version 1.2.2), LightGBM (version 3.3.4), XGBoost (version 1.7.5), optuna (version 3.1.1). The version of ChatGPT utilized in this study is gpt-3.5-turbo-0301. Relevant codes and sample data will be offered in the final version.

The following experiments will validate the performance of GPT-LGBM (Path-D) and GPT-LGBM (Path-K), where GPT-LGBM (Path-D) refers to using GPT-LGBM along Path-D and GPT-LGBM (Path-K) refers to using GPT-LGBM via Path-K.

4.3 Experiments of GPT-LGBM (Path-D)

4.3.1 Comparison with benchmark models

To validate the effectiveness of the proposed GPT-LGBM (Path-D) model, extensive experiments are conducted to compare with the 17 benchmark models in terms of AUC, KS, F1 Score and G-mean. The experimental results are presented in Table 1 which reveals that GPT-LGBM (Path-D) exhibits superior performance against the other models.

More specifically, GPT-LGBM (Path-D) achieves 71.9404% AUC, 36.9754% KS, 39.1636 % F1 Score and 68.0837% G-mean. It has the best performance on KS, AUC and G-mean, while obtaining the second-best performance on F1 Score. In terms of the F1 Score, the performance gap between GPT-LGBM (Path-D) and the best model is only 0.0031%. Utilizing the Wilcoxon Signed-Rank test, GPT-LGBM (Path-D) exhibits overall statistically significant superiority over the remaining comparative models. Furthermore, GPT-LGBM (Path-D) has a probability of over 95% of being superior to every benchmark model in both the KS and G-mean metrics.

In addition, Table 1 reveals two important findings. First, comparing the models that utilize only structured features to those that combine structured features and psychological features, it is evident that the latter exhibits better performance. For example, the incorporation of Big Five personality traits improves AUC by 0.3291% for LR, 0.8365%

for SVM. These comparisons demonstrate that the integration of psychological features obtained through ChatGPT contributes to the improved performance of loan default prediction. Second, for the models that solely rely on the structured features, LightGBM performs best on F1 Score, and second-best on AUC, KS and G-mean. It indicates that LightGBM is a powerful classifier for credit assessment.

Table 1
The performance comparison of the benchmark models and GPT-LGBM (Path-D). Big Five (Path-D) refers to Big Five personality traits extracted through Path-D. The highest value for each performance metric is shown in bold. The markers *, **, and *** refer to the Wilcoxon Signed-Rank test p-value <0.1, 0.05, and 0.01 respectively when comparing with GPT-LGBM (Path-D). The number in each bracket is the gain of GPT-LGBM (Path-D) compared to the benchmark model.

Features	Model	AUC (%)	KS (%)	F1 (%)	G-mean (%)
Structured features	LR	71.3602*** (0.5801)	35.5572** (1.4182)	38.8270 (0.3366)	67.4035** (0.6802)
	KNN	65.8398*** (6.1006)	25.8710*** (11.1044)	34.1355*** (5.0281)	61.7299*** (6.3538)
	SVM	63.6616*** (8.2788)	23.8298*** (13.1456)	33.2452*** (5.9184)	60.6870*** (7.3967)
	MLP	71.5323** (0.4080)	35.9738** (1.0016)	38.7847 (0.3789)	67.5840** (0.4996)
	DT	66.8511*** (5.0893)	28.2422*** (8.7332)	34.9698*** (4.1938)	63.1266*** (4.9571)
	RF	71.2609*** (0.6794)	34.8349*** (2.1405)	38.3798*** (0.7838)	67.0772*** (1.0064)
	GBDT	71.4804*** (0.4599)	35.3711*** (1.6044)	38.7865* (0.3771)	67.1980*** (0.8856)
	XGBoost	71.1972*** (0.7432)	35.4773*** (1.4982)	39.1604 (0.0032)	67.4671*** (0.6165)
	LightGBM	71.5309*** (0.4094)	35.8357*** (1.1397)	**39.1667** (-0.0031)	67.5135*** (0.5702)
Structured features + Big Five (Path-D)	LR	71.6893* (0.2510)	35.7838** (1.1916)	38.8027 (0.3609)	67.4718** (0.6119)
	KNN	65.9284*** (6.0120)	26.2178*** (10.7577)	34.0434*** (5.1202)	61.9464*** (6.1373)
	SVM	64.4981*** (7.4422)	25.2514*** (11.7240)	33.6623*** (5.5013)	61.0964*** (6.9873)
	MLP	71.6785 (0.2618)	35.3912** (1.5842)	38.3303** (0.8333)	67.1159*** (0.9678)
	DT	67.3015*** (4.6388)	28.8250*** (8.1504)	35.3852*** (3.7784)	63.4990*** (4.5846)
	RF	71.6028* (0.3376)	34.8741*** (2.1013)	38.4401*** (0.7235)	67.0390*** (1.0447)
	GBDT	71.8525 (0.0878)	35.7178*** (1.2577)	39.0148 (0.1487)	67.3385*** (0.7451)
	XGBoost	71.5933** (0.3471)	35.5284*** (1.4470)	38.8476 (0.3160)	67.2974*** (0.7862)
	GPT-LGBM (Path-D)	**71.9404** (/)	**36.9754** (/)	39.1636 (/)	**68.0837** (/)

4.3.2 Ablation Study

To verify the ability of ChatGPT to extract credit-relevant information from textual data, ablation experiments are conducted with the same settings in previous experiments.

Firstly, GPT-LGBM (Path-D) is compared with two similar models that employ RoBERTa, which is a state-of-the-art pre-trained language model for credit evaluation [12]. One, denoted as RoBERTa-LGBM, utilizes RoBERTa for text feature extraction instead of ChatGPT, while following the rest procedures of GPT-LGBM (Path-D). That is, each 768-dimensional vector generated by RoBERTa and the corresponding structured features are concatenated and fed into LightGBM for classification. The other, denoted as RoBERTa-PCA-LGBM, differs from RoBERTa-LGBM in that it employs PCA after RoBERTa. Specifically, PCA is used as an unsupervised technique to transform the outputs of RoBERTa into 5-dimensional vectors. In this way, RoBERTa-PCA-LGBM can generate textual features of the same dimension as GPT-LGBM (Path-D), except for its lack of psychological meaning.

As illustrated in Table 2, GPT-LGBM (Path-D) outperforms both RoBERTa-LGBM and RoBERTa-PCA-LGBM across all four metrics. In the Wilcoxon Signed-Rank test, GPT-LGBM (Path-D) passed the 5% significance test, particularly showing significant results at the 1% level in terms of KS and AUC. It confirms the significance of the Big Five personality traits retrieved by ChatGPT for loan default prediction.

Secondly, GPT-LGBM (Path-D) is compared with another similar model named Chat-LGBM-TextOnly, which removes the structured features from the inputs of LightGBM. Without structured data, Chat-LGBM-TextOnly exhibits significantly poorer performance with 10%, as illustrated in Table 2. It reveals that using the psychological features alone is inadequate, underscoring the importance of combining textual and structured data for default prediction.

Table 2
Ablation experiments. This table reports the results when the input features are changed. The markers *, **, and *** refer to the Wilcoxon Signed-Rank test p-value <0.1, 0.05, and 0.01 respectively when comparing with GPT-LGBM (Path-D). The number in each bracket is the gain of GPT-LGBM (Path-D) compared to the three models.

Model	AUC (%)	KS (%)	F1 (%)	G-mean (%)
Chat-LGBM-TextOnly	58.2130*** (13.7273)	15.7057*** (21.2698)	28.9263*** (10.2373)	55.4683*** (12.6154)
RoBERTa-LGBM	71.4157* (0.5247)	35.2954*** (1.6800)	38.4916** (0.6720)	67.1778*** (0.9059)
RoBERTa-PCA-LGBM	71.5492* (0.3911)	35.2214*** (1.7540)	38.3809** (0.7827)	67.0433*** (1.0403)

4.3.3 Feature importance analysis

For credit scoring, interpretability is very important since it can help risk managers better understand the model and the impact of each feature. After GPT-LGBM (Path-D) transforms the credit-related texts into five psychological features, the feature importance of both psychological features and structured features can be obtained by counting the number of times each feature is used in LightGBM. As shown in Fig. 4, 'purpose', 'sub_grade', 'annual_inc, 'inst_to_inc' and 'hpi_yoychg' are the top five features that come from the structured features. The psychological features are positioned in the middle tier. Although the psychological features are not as important as some structured features, they still play a significant role in loan default prediction. The most important feature among the psychological features is 'Extraversion', ranking at the 13th position, with 'Agreeableness' and 'Conscientiousness' subsequently ranked the 14th and 16th. Interestingly, 'Openness' is the least important feature among the psychological features.

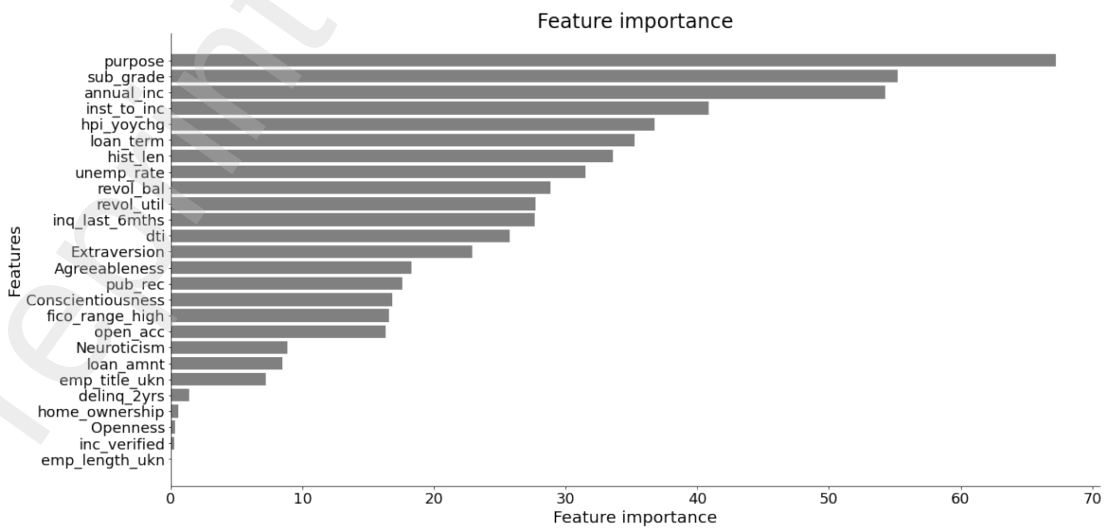

Fig. 4. Feature importance for loan default prediction

4.4 Experiments of GPT-LGBM (Path-K)

To evaluate the performance of GPT-LGBM along Path-K, i.e. GPT-LGBM (Path-K), ten round experiments on the knowledge distillation model are conducted, with each round using a predetermined amount of training data.

Initially the whole dataset is split into a training dataset and a test dataset with a ratio of 7:3. Then 10 knowledge distillation models in Table 3 are respectively trained with a varying portion of the training dataset, ranging from 10% up to 100% with a step of 10%. Each trained knowledge distillation model is used to predict Big Five personality traits based on the loan texts in the test dataset. These predicted values are used as textual features, which are combined with the structured features and loan statuses in the test dataset to generate a new dataset. That is, each new dataset has the same structured features and loan statuses but different textual features from different knowledge distillation models. Consequently, LightGBM is trained and tested on each new dataset using the same experimental settings as in Section 4.2. These procedures yield the prediction results of the 10 different GPT-LGBM (Path-K) models.

As presented in Table 3, the prediction results of the 10 GPT-LGBM (Path-K) models are generally superior to the LightGBM model that is solely based on structured features. Specifically, the maximum values of AUC, KS, F1 Score and G-mean for the GPT-LGBM (Path-K) models reached 68.8858%, 35.3791%, 39.2263%, and 66.6418% respectively. Hence, when only partial data can be accessed by ChatGPT, employing the GPT-LGBM (Path-K) model becomes a preferred option for credit scoring. However, compared to the results obtained by the GPT-LGBM (Path-D) model with the same datasets, the best-performing GPT-LGBM (Path-K) models exhibit a -0.3893%, -0.6456%, 0.2373%, and -0.3920% gap w.r.t. AUC, KS, F1 Score and G-mean. It indicates that for the loan texts, the feature extraction ability of ChatGPT is better than the knowledge distillation model with RoBERTa. As the percentage of data used for training the knowledge distillation model increases, the performance of the GPT-LGBM (Path-K) model gradually improves. Interestingly, the GPT-LGBM (Path-K) model does not always reach its best performance at the largest dataset size (100%). This phenomenon may be attributed to the randomness of the deep learning algorithm in the knowledge distillation model.

Table 3
Experimental results of GPT-LGBM (Path-K) based on different proportions of data. Big Five (Path-K) refers to Big Five personality traits extracted through Path-K. GPT-LGBM (Path-K)-X% refers to GPT-LGBM (Path-K) of which knowledge distillation model is trained on X% of the training dataset. The highest value for each performance metric is shown in bold.

Features	Model	AUC (%)	KS (%)	F1 (%)	G-mean (%)
Structured features	LightGBM	67.9594	33.8690	38.1843	65.6645
Structured features + Big Five (Path-K)	GPT-LGBM (Path-K)-10%	68.3297	33.6941	38.2634	65.4086
	GPT-LGBM (Path-K)-20%	68.1801	34.0884	37.8003	65.7121
	GPT-LGBM (Path-K)-30%	68.7060	35.2558	**39.2263**	66.0992
	GPT-LGBM (Path-K)-40%	68.5647	34.0842	38.0796	65.6802
	GPT-LGBM (Path-K)-50%	68.5116	34.7244	38.7637	66.4082
	GPT-LGBM (Path-K)-60%	68.8858	34.8522	38.3444	66.6418
	GPT-LGBM (Path-K)-70%	68.6277	35.3398	38.5244	66.0832
	GPT-LGBM (Path-K)-80%	68.6444	34.5324	38.0918	66.5749
	GPT-LGBM (Path-K)-90%	68.4110	34.3802	38.4701	66.1205
	GPT-LGBM (Path-K)-100%	68.2418	35.3791	38.8092	66.4751
Structured features + Big Five (Path-D)	GPT-LGBM (Path-D)	**69.2751**	**36.0247**	38.9890	**67.0338**

4.5 Further experiments with ChatGPT

In order to further investigate the capability of ChatGPT in the field of credit assessment, two ChatGPT-only models are constructed. In these models, new prompts are designed according to the method mentioned in Section 3.1 with the difference lying in using ChatGPT directly for loan default prediction.

In the first model denoted as GPT-text, only the loan texts are transformed into new prompts, and ChatGPT is employed to determine the applicants' risk of default with these prompts. The second model, denoted as GPT-text&str, involves converting structured data into textual format and merging it with the loan texts, followed by generating the corresponding new prompts and feeding them into ChatGPT for measuring the applicants' risk. Subsequently, experiments are performed on both models with similar experimental settings as in Section 4.2

As shown in Table 4, GPT-text&str outperforms GPT-text in all performance metrics by an approximate margin of 4%, which highlights the importance of structured data for credit scoring. However, compared with the classifiers such as GBDT, XGBoost, LightGBM, the performance of GPT-text&str is significantly lower when using the same datasets. The above results suggest that using only ChatGPT to deal with structured and textual data may not be the most effective credit-scoring approach.

Table 4
Experimental results with different input texts using ChatGPT directly. That is, ChatGPT is the only model used for loan default prediction.

Model	AUC (%)	KS (%)	F1 (%)	G-mean (%)
GPT-text	57.1564	14.3447	27.9723	54.4366
GPT-text&str	61.4855	20.0086	31.2372	58.7946

5. Conclusion

This study introduces an integrated framework GPT-LGBM to incorporate both structured data and loan texts for credit scoring. Since ChatGPT excels at reasoning and inference tasks, this artificial intelligence language model is used to transform credit-related textual data into Big Five personality traits without manual annotation. The resulted Big Five personality traits, along with structured features, are subsequently inputted into the LightGBM classifier to predict loan defaults. Due to different data privacy requirements, two paths are developed, with one entailing direct textual feature extraction and the other combining ChatGPT and RoBERTa based on knowledge distillation. Experimental results on the real-world data from Lending Club demonstrate that GPT-LGBM can achieve superior performance for credit evaluation while maintaining interpretability of features.

Further works can be done to improve the proposed framework. The first is to incorporate other large language models such as GPT-4 and LLaMA for textual feature extraction. The second is to validate the framework on more credit-related datasets. The third is to implement the proposed framework in other areas whenever both structured and textual data are available.

Appendix A

Table A.1 The hyperparameter search spaces

Model	Hyperparameters
LR	solver $\in \{'lbfgs','liblinear','newton-cg','newton-cholesky','sag','saga'\}$
	penalty $\in$
	$\{'l2',None \mid solver \in \{'lbfgs','newton-cg','newton-cholesky','sag'\}\}$
	$\cup \{'l1','l2' \mid solver \in \{'liblinear'\}\}$
	$\cup \{'elasticnet','l1','l2',None \mid solver \in \{'saga'\}\}$
	C $\in \{1e-4, 1e-3, 1e-2, 1e-1, 1, 10, 100, 1000, 10000\}$
KNN	n_neighbors $\in \{3, 4, \ldots, 19, 20\}$
	weights $\in \{'uniform', 'distance'\}$
	leaf_size $\in \{1, 5, 10, 20, \ldots, 90, 100\}$
	p $\in \{1, 2\}$
SVM	C $\in \{1e-4, 1e-3, 1e-2, 1e-1, 1, 10, 100\}$
	kernel $\in \{'linear','poly','rbf','sigmoid'\}$
	degree $\in \{2, 3, 4, 5\}$
	gamma $\in \{'scale','auto'\}$
	coef0 $\in \{0, 0.1, \ldots, 0.9, 1.0\}$
MLP	n_layers $\in \{1, 2\}$
	hidden_layer_size $\in$
	$\{(x,) \mid n\_layers = 1, x \in [feature\_cnt, 2 * feature\_cnt]\}$
	$\cup \{(x, y) \mid n\_layers = 2, x, y \in [feature\_cnt, 2 * feature\_cnt]\}$
	activation $\in \{'identity','logistic','tanh','relu'\}$
	solver $\in \{'lbfgs','sgd','adam'\}$
	alpha $\in \{1e-5, 1e-4, 1e-3, 1e-2, 1e-1, 1, 10\}$
	batch_size $\in \{64, 128, \ldots, 1024, 2048\}$
	learning_rate $\in \{'constant','invscaling','adaptive'\}$
	learning_rate_init $\in \{1e-4, 1e-3, 1e-2, 1e-1\}$
DT	criterion $\in \{'gini', 'entropy', 'log\_loss'\}$
	max_depth $\in \{3, 4, \ldots, 19, 20\}$
	min_samples_leaf $\in \{1, 5, 10, 20, 50\}$
	min_samples_split $\in \{1, 5, \ldots, \text{min\_samples\_leaf} * 2\}$
	max_features $\in \{'sqrt', 'log2', None\}$
RF	n_estimators $\in \{100, 150, \ldots, 950, 1000\}$
	criterion $\in \{'gini', 'entropy','log\_loss'\}$
	max_depth $\in \{3, 4, \ldots, 19, 20\}$
	min_samples_leaf $\in \{1, 5, 10, 20, 50\}$
	min_samples_split $\in \{1, 5, \ldots, \text{min\_samples\_leaf} * 2\}$
	max_features $\in \{'sqrt', 'log2', None\}$
GBDT	learning_rate $\in \{0.01, 0.05, 0.1\}$
	n_estimators $\in \{100, 150, \ldots, 550, 600\}$
	subsample $\in \{0.1, 0.2, \ldots, 0.9, 1.0\}$
	min_samples_leaf $\in \{1, 5, 10, 20, 50\}$
	min_samples_split $\in \{1, 5, \ldots, \text{min\_samples\_leaf} * 2\}$
	max_depth $\in \{3, 4, 5, 6, 7, 8\}$
	max_features $\in \{'sqrt','log2', None\}$
XGBoost	learning_rate $\in \{0.1, 0.05, 0.01\}$
	n_estimators $\in \{100, 150, \ldots, 550, 600\}$

	max_depth $\in \{3, 4, \ldots, 7, 8\}$
	gamma $\in \{1e-4, 1e-3, 1e-2, 0.1, 0.2, \ldots, 0.9, 1.0\}$
	subsample $\in \{0.1, 0.2, \ldots, 0.9, 1.0\}$
	colsample_bytree $\in \{0.1, 0.2, \ldots, 0.9, 1.0\}$
	reg_alpha $\in \{0, 0.1, \ldots, 0.9, 1\}$
	reg_lambda $\in \{0, 0.1, \ldots, 0.9, 1\}$
	tree_method $\in \{'exact'\}$
LightGBM	learning_rate $\in \{0.1, 0.05, 0.01\}$
	max_depth $\in \{3, 4, \ldots, 7, 8\}$
	num_leaves $\in \{3,5,7,\ldots,2^{max\_depth} - 1\}$
	subsample $\in \{0.5, 0.6, \ldots, 0.9, 1.0\}$
	colsample_bytree $\in \{0.5, 0.6, \ldots, 0.9, 1.0\}$
	reg_alpha $\in \{0, 0.1, \ldots, 0.9, 1\}$
	reg_lambda $\in \{0, 0.1, \ldots, 0.9, 1\}$
	min_child_samples $\in \{20\}$
	early_stopping_rounds $\in \{0\}$
	n_estimators $\in \{100, 150, \ldots, 550, 600\}$
	categorical_feature $\in \{categorical\ features\}$

References

[1] D.J. Hand, W.E. Henley, Statistical Classification Methods in Consumer Credit Scoring: a Review, Journal of the Royal Statistical Society: Series A (Statistics in Society). 160 (1997) 523–541. https://doi.org/10.1111/j.1467-985X.1997.00078.x.

[2] H.A. Abdou, J. Pointon, Credit Scoring, Statistical Techniques and Evaluation Criteria: A Review of the Literature, Intelligent Systems in Accounting, Finance and Management. 18 (2011) 59–88. https://doi.org/10.1002/isaf.325.

[3] C. Jiang, Z. Wang, H. Zhao, A prediction-driven mixture cure model and its application in credit scoring, European Journal of Operational Research. 277 (2019) 20–31. https://doi.org/10.1016/j.ejor.2019.01.072.

[4] J. Abellán, J.G. Castellano, A comparative study on base classifiers in ensemble methods for credit scoring, Expert Systems with Applications. 73 (2017) 1–10. https://doi.org/10.1016/j.eswa.2016.12.020.

[5] Z. Wang, C. Jiang, H. Zhao, Y. Ding, Mining Semantic Soft Factors for Credit Risk Evaluation in Peer-to-Peer Lending, Journal of Management Information Systems. 37 (2020) 282–308. https://doi.org/10.1080/07421222.2019.1705513.

[6] W. Zhang, C. Wang, Y. Zhang, J. Wang, Credit risk evaluation model with textual features from loan descriptions for P2P lending, Electronic Commerce Research and Applications. 42 (2020) 100989. https://doi.org/10.1016/j.elerap.2020.100989.

[7] K. Yang, H. Yuan, R.Y.K. Lau, PsyCredit: An interpretable deep learning-based credit assessment approach facilitated by psychometric natural language processing, Expert Systems with Applications. 198 (2022) 116847. https://doi.org/10.1016/j.eswa.2022.116847.

[8] G. Guo, F. Zhu, E. Chen, Q. Liu, L. Wu, C. Guan, From Footprint to Evidence: An Exploratory Study of Mining Social Data for Credit Scoring, ACM Trans. Web. 10 (2016) 22:1-22:38. https://doi.org/10.1145/2996465.

[9] T. Lu, Y. Xu, G. Chen, C. Zhang, Your Posts Expose You: Theory-Driven Approach to Credit Risk Prediction for Microloans Based on Social Media Content, (2022). https://doi.org/10.2139/ssrn.4138565.

[10] Y. Wang, W. Xu, Leveraging deep learning with LDA-based text analytics to detect automobile insurance fraud, Decision Support Systems. 105 (2018) 87–95. https://doi.org/10.1016/j.dss.2017.11.001.

[11] Y. Xia, L. He, Y. Li, N. Liu, Y. Ding, Predicting loan default in peer-to-peer lending using narrative data, Journal of Forecasting. 39 (2020) 260–280. https://doi.org/10.1002/for.2625.

[12] J. Kriebel, L. Stitz, Credit default prediction from user-generated text in peer-to-peer lending using deep learning, European Journal of Operational Research. 302 (2022) 309–323. https://doi.org/10.1016/j.ejor.2021.12.024.

[13] Z. Li, H. Zhang, M. Yu, H. Wang, Too long to be true in the description? Evidence from a Peer-to-Peer platform in China, Research in International Business and Finance. 50 (2019) 246–251. https://doi.org/10.1016/j.ribaf.2019.06.005.

[14] C. Jiang, Z. Wang, R. Wang, Y. Ding, Loan default prediction by combining soft information extracted from descriptive text in online peer-to-peer lending, Ann Oper Res. 266 (2018) 511–529. https://doi.org/10.1007/s10479-017-2668-z.

[15] E.A.M. van Dis, J. Bollen, W. Zuidema, R. van Rooij, C.L. Bockting, ChatGPT: five priorities for research, Nature. 614 (2023) 224–226. https://doi.org/10.1038/d41586-023-00288-7.

[16] M. Dowling, B. Lucey, ChatGPT for (Finance) research: The Bananarama Conjecture, Finance Research Letters. 53 (2023) 103662. https://doi.org/10.1016/j.frl.2023.103662.

[17] F. Fui-Hoon Nah, R. Zheng, J. Cai, K. Siau, L. Chen, Generative AI and ChatGPT: Applications, challenges, and AI-human collaboration, Journal of Information Technology Case and Application Research. 25 (2023) 277–304. https://doi.org/10.1080/15228053.2023.2233814.

[18] Q. Zhong, L. Ding, J. Liu, B. Du, D. Tao, Can ChatGPT Understand Too? A Comparative Study on ChatGPT and Fine-tuned BERT, (2023). https://doi.org/10.48550/arXiv.2302.10198.

[19] C. Qin, A. Zhang, Z. Zhang, J. Chen, M. Yasunaga, D. Yang, Is ChatGPT a General-Purpose Natural Language Processing Task Solver?, (2023). https://doi.org/10.48550/arXiv.2302.06476.

[20] M.M. Amin, E. Cambria, B.W. Schuller, Will Affective Computing Emerge From Foundation Models and General Artificial Intelligence? A First Evaluation of ChatGPT, IEEE Intelligent Systems. 38 (2023) 15–23. https://doi.org/10.1109/MIS.2023.3254179.
[21] H. Rao, C. Leung, C. Miao, Can ChatGPT Assess Human Personalities? A General Evaluation Framework, (2023). https://doi.org/10.48550/arXiv.2303.01248.
[22] Z. Wang, J. Xiao, L. Wang, J. Yao, A novel federated learning approach with knowledge transfer for credit scoring, Decision Support Systems. (2023) 114084. https://doi.org/10.1016/j.dss.2023.114084.
[23] H. He, Z. Wang, H. Jain, C. Jiang, S. Yang, A privacy-preserving decentralized credit scoring method based on multi-party information, Decision Support Systems. 166 (2023) 113910. https://doi.org/10.1016/j.dss.2022.113910.
[24] M. Gurman, Samsung bans staff's AI use after spotting ChatGPT data leak, The Japan Times. (2023). https://www.japantimes.co.jp/news/2023/05/02/business/tech/samsung-bans-chatgpt-workplace-use/ (accessed December 8, 2023).
[25] T.-S. Lee, C.-C. Chiu, Y.-C. Chou, C.-J. Lu, Mining the customer credit using classification and regression tree and multivariate adaptive regression splines, Computational Statistics & Data Analysis. 50 (2006) 1113–1130. https://doi.org/10.1016/j.csda.2004.11.006.
[26] I.-C. Yeh, C. Lien, The comparisons of data mining techniques for the predictive accuracy of probability of default of credit card clients, Expert Systems with Applications. 36 (2009) 2473–2480. https://doi.org/10.1016/j.eswa.2007.12.020.
[27] S. Bhattacharyya, S. Jha, K. Tharakunnel, J.C. Westland, Data mining for credit card fraud: A comparative study, Decision Support Systems. 50 (2011) 602–613. https://doi.org/10.1016/j.dss.2010.08.008.
[28] S. Lessmann, B. Baesens, H.-V. Seow, L.C. Thomas, Benchmarking state-of-the-art classification algorithms for credit scoring: An update of research, European Journal of Operational Research. 247 (2015) 124–136. https://doi.org/10.1016/j.ejor.2015.05.030.
[29] Y. Xia, C. Liu, Y. Li, N. Liu, A boosted decision tree approach using Bayesian hyper-parameter optimization for credit scoring, Expert Systems with Applications. 78 (2017) 225–241. https://doi.org/10.1016/j.eswa.2017.02.017.
[30] Y.-C. Chang, K.-H. Chang, G.-J. Wu, Application of eXtreme gradient boosting trees in the construction of credit risk assessment models for financial institutions, Applied Soft Computing. 73 (2018) 914–920. https://doi.org/10.1016/j.asoc.2018.09.029.
[31] H. He, W. Zhang, S. Zhang, A novel ensemble method for credit scoring: Adaption of different imbalance ratios, Expert Systems with Applications. 98 (2018) 105–117. https://doi.org/10.1016/j.eswa.2018.01.012.
[32] Y. Xia, C. Liu, B. Da, F. Xie, A novel heterogeneous ensemble credit scoring model based on bstacking approach, Expert Systems with Applications. 93 (2018) 182–199. https://doi.org/10.1016/j.eswa.2017.10.022.
[33] W. Liu, H. Fan, M. Xia, M. Xia, A focal-aware cost-sensitive boosted tree for imbalanced credit scoring, Expert Systems with Applications. 208 (2022) 118158. https://doi.org/10.1016/j.eswa.2022.118158.
[34] G. Dorfleitner, C. Priberny, S. Schuster, J. Stoiber, M. Weber, I. de Castro, J. Kammler, Description-text related soft information in peer-to-peer lending – Evidence from two leading European platforms, Journal of Banking & Finance. 64 (2016) 169–187. https://doi.org/10.1016/j.jbankfin.2015.11.009.
[35] B.-H. Nguyen, V.-N. Huynh, Textual analysis and corporate bankruptcy: A financial dictionary-based sentiment approach, Journal of the Operational Research Society. 73 (2022) 102–121. https://doi.org/10.1080/01605682.2020.1784049.
[36] H. Liu, M. Yuan, M. Zhou, How Does the Urgency of Borrowing in Text Messages Affect Loan Defaults? Evidence from P2P Loans in China, Security and Communication Networks. 2021 (2021) e4060676. https://doi.org/10.1155/2021/4060676.
[37] F. Mai, S. Tian, C. Lee, L. Ma, Deep learning models for bankruptcy prediction using textual disclosures, European Journal of Operational Research. 274 (2019) 743–758. https://doi.org/10.1016/j.ejor.2018.10.024.
[38] R. Matin, C. Hansen, C. Hansen, P. Mølgaard, Predicting distresses using deep learning of text segments in annual reports, Expert Systems with Applications. 132 (2019) 199–208. https://doi.org/10.1016/j.eswa.2019.04.071.
[39] M. Stevenson, C. Mues, C. Bravo, The value of text for small business default prediction: A Deep Learning approach, European Journal of Operational Research. 295 (2021) 758–771. https://doi.org/10.1016/j.ejor.2021.03.008.

[40] B. Huang, X. Yao, Y. Luo, J. Li, Improving financial distress prediction using textual sentiment of annual reports, Ann Oper Res. 330 (2023) 457–484. https://doi.org/10.1007/s10479-022-04633-3.
[41] Y. Bang, S. Cahyawijaya, N. Lee, W. Dai, D. Su, B. Wilie, H. Lovenia, Z. Ji, T. Yu, W. Chung, Q.V. Do, Y. Xu, P. Fung, A Multitask, Multilingual, Multimodal Evaluation of ChatGPT on Reasoning, Hallucination, and Interactivity, (2023). https://doi.org/10.48550/arXiv.2302.04023.
[42] J. White, Q. Fu, S. Hays, M. Sandborn, C. Olea, H. Gilbert, A. Elnashar, J. Spencer-Smith, D.C. Schmidt, A Prompt Pattern Catalog to Enhance Prompt Engineering with ChatGPT, (2023). https://doi.org/10.48550/arXiv.2302.11382.
[43] Z. Li, Y. Chen, X. Zhang, X. Liang, BookGPT: A General Framework for Book Recommendation Empowered by Large Language Model, Electronics. 12 (2023) 4654. https://doi.org/10.3390/electronics12224654.
[44] ChatGPT Prompt Engineering for Developers, (n.d.). https://learn.deeplearning.ai/chatgpt-prompt-eng/lesson/1/introduction (accessed December 8, 2023).
[45] J. Gou, B. Yu, S.J. Maybank, D. Tao, Knowledge Distillation: A Survey, Int J Comput Vis. 129 (2021) 1789–1819. https://doi.org/10.1007/s11263-021-01453-z.
[46] F.M. Thoker, J. Gall, Cross-Modal Knowledge Distillation for Action Recognition, in: 2019 IEEE International Conference on Image Processing (ICIP), 2019: pp. 6–10. https://doi.org/10.1109/ICIP.2019.8802909.
[47] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, V. Stoyanov, RoBERTa: A Robustly Optimized BERT Pretraining Approach, (2019). https://doi.org/10.48550/arXiv.1907.11692.
[48] G. Ke, Q. Meng, T. Finley, T. Wang, W. Chen, W. Ma, Q. Ye, T.-Y. Liu, LightGBM: A Highly Efficient Gradient Boosting Decision Tree, in: Advances in Neural Information Processing Systems, Curran Associates, Inc., 2017. https://proceedings.neurips.cc/paper/2017/hash/6449f44a102fde848669bdd9eb6b76fa-Abstract.html (accessed December 8, 2023).
[49] T. Fitzpatrick, C. Mues, How can lenders prosper? Comparing machine learning approaches to identify profitable peer-to-peer loan investments, European Journal of Operational Research. 294 (2021) 711–722. https://doi.org/10.1016/j.ejor.2021.01.047.
[50] J. Wu, X.-Y. Chen, H. Zhang, L.-D. Xiong, H. Lei, S.-H. Deng, Hyperparameter Optimization for Machine Learning Models Based on Bayesian Optimizationb, Journal of Electronic Science and Technology. 17 (2019) 26–40. https://doi.org/10.11989/JEST.1674-862X.80904120.