第 6 回とめ研究所若手研究者懸賞論文

KGSynX: 知識グラフと説明可能なフィードバックによる LLM の合成表形式データ生成

Yu Ke[1], 石倉茂[2], 至極有輝[2], 臼倉由香利[2], 早矢仕晃章[1]

1 東京大学 大学院工学系研究科 システム創成学専攻
2 株式会社インフォマート

1. はじめに

1.1 背景

データは 21 世紀の石油と形容されているように、人工知能技術とともにイノベーションの新しい源泉として注目されている。このような中、近年データの重要性はますます高まっており、企業はデータを交換可能な商材として取引し始めている [1]。さらに、分散型機械学習技術の普及に伴い、エージェント同士がデータを共有できるデータ取引市場の必要性が高まってきている[2]。

しかし、価値のあるデータセットにはしばしばユーザーのプライバシーや組織の機密情報が含まれるため、倫理的・法的な課題が内在する。そのため、研究コミュニティや産業界では代替手段として合成データ（synthetic data）が注目を集めている [3][4]。合成データとは、実際の個人情報を含まずに現実世界の特徴を模倣した人工的に生成されたデータのことである。合成データは、プライバシー保護、研究用データの可用性向上、機械学習モデルにおけるバイアスの低減といったメリットがあり [5]、信頼される AI 開発において必要不可欠な技術として期待が高まっている。実際、ガートナー[1]は 2030 年までに AI モデルで利用される合成データの量が実データを上回ると予測している（図 1）。また、市場調査レポート[2]によれば、合成データ生成市場は 2024 年の 3 億 1,350 万 USD から 2034 年には約 66 億 3,798 万 USD に達し、予測期間中に年平均成長率 35.7%という爆発的な成長を遂げると予測している（図 2）。

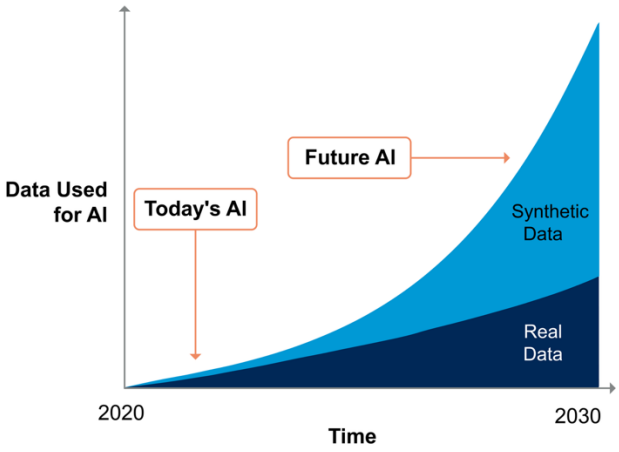

Figure 1: The trend of synthetic data

[1] https://gretel.ai/technical-glossary/what-is-synthetic-data

[2] https://market.us/report/synthetic-data-generation-market/

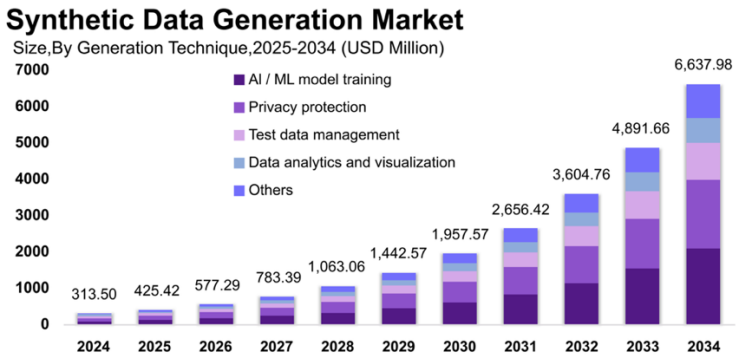

Figure 2: The value of synthetic data market

現代のデータ中心社会において、表形式データは臨床診断や金融リスクモデリングから顧客分析、物流最適化に至るまで、幅広い分野の意思決定の基盤として機能し続けている [6]。しかし、EU の一般データ保護規則（GDPR）や米国の医療保険の携行性・責任に関する法律（HIPAA）などの規制により、こうしたデータはしばしば組織内に死蔵され、利用が制限されている [7]。匿名化手法 [8] を用いても、高次元データにおいては再識別防止やデータの有用性維持が十分でない場合がある。そのため、実データの重要な統計的・構造的特性を保持しつつ、機微な情報を含まない人工データを生成する合成データ生成手法が求められている [9]。

一方で、合成データは既に実用化されたケースが存在する。例えば、英国国民保健サービス（NHS）は研究用に合成 EHR を公開する取り組みを進めている [10]。金融分野では、合成クレジットカード取引データが不正検知モデルのシミュレーションに活用されている [11]。さらに、SAP や Snowflake といった企業は、ソフトウェアの内部テストやデータ統合の検証を目的に合成データを利用している [12]。

しかし、深層生成モデルによって画像・動画・自然言語の合成が進展している一方で、表形式データの合成は依然として困難である [13]。その理由は、表形式データがカテゴリ、順序性、連続性といった多様な特徴を含み、変数（カラム）間に複雑な関係を持っているためである。例えば通信会社の顧客離反データでは、通話回数、契約タイプ、課金プランなどの変数が相互依存しており、医療データでは検査値や年齢が診断結果と整合していなければならない。

従来のアプローチは、SMOTE によるオーバーサンプリング、ブートストラップ法によるデータ拡張、ガウス・コピュラなどの統計分布に基づくサンプリングなどが利用されてきた。しかし、これらは特徴量ごとに処理されることが多く、高次の依存関係やドメイン特有の制約を反映できない [14]。

合成されたデータの評価についても課題がある。既存の指標（特徴量相関、KL ダイバージェンス、分類精度など）は、生成されたデータの分布や表面的な性能を測ることはできるが、「合成データが実データで訓練したモデルと同じように振る舞うか」ということを測ることはできない。特に重要なのは、特徴量同士の相互作用によって出力がどのように決定されるかという意味的な関係性が保持されているかどうかである。この点を捉えられないため、合成データが分布上は実データに近くても、まったく異なる論理を持つモデルを生み出してしまう危険がある。その結果、偏った分析や誤った判断につながる可能性がある。

こうした課題に対し、解釈可能な機械学習の進展が新しい方法を提供している。特に、SHAP（SHapley Additive exPlanations） [15]は、特徴量が予測に与える寄与を定量化できるモデル非依存の手法である。SHAP 値を用いることで、合成データと実データで学習したモデルの意思決定プロセスを比較でき、差異に基づいて生成プロセスを改善するフィードバックループを構築することが可能になる。

合成表形式データは、プライバシー保護とデータ効率を両立する有望な手段である。しかし、それが真に上手く機能するためには、単なる統計的な分布の模倣ではなく、意味的・構造的な一貫性をどのように維持できるかにかかっている。本研究は、知識グラフ、LLM（Large Language Model）、SHAP に基づく評価を統合し、知識駆動の制御と説明可能なフィードバックを組み合わせた新しい合成データ生成手法を提案する。

1.2 研究目的とアプローチ

本論文の目的は、実世界のデータセットが持つ統計的特性、構造的な依存関係、そして意思決定に関わる意味を忠実に保ったまま、表形式の合成データを生成するための、原理に裏づけられたフレームワークを開発することである。従来の生成モデルが分布上の類似性に主眼を置いてきたのに対し、本手法は意味的な整合性と構造的な制御を重視し、合成データを下流の分析や予測タスクにおいて、合成データが「実データの信頼できる代替データ」として機能させることを目指す。

この目的達成のために、本研究では新しい合成データ生成手法 KGSynX を提案する。KGSynX は、主に以下の 3 つを統合したハイブリッド・フレームワークである。

1. 知識グラフに基づく構造モデリング [16]
2. LLM によるデータ生成 [17]
3. SHAP による帰属情報を用いたフィードバックによる改善 [15]

このコアとなるアイデアは、表形式データの関係構造や論理制約を中間表現として知識グラフ化し、その構造情報を保持したプロンプトを LLM に渡せば、意味的一貫性と統計的な現実性を兼ね備えた合成データを生成できる、という点にある。さらに、SHAP に基づくフィードバックを取り入れることで、実データ学習モデルと合成データ学習モデルの間に生じる意味的な乖離を反復的に縮小できると考えた。図 3 にて、本研究の生成手法と従来手法のパイプラインを示す。

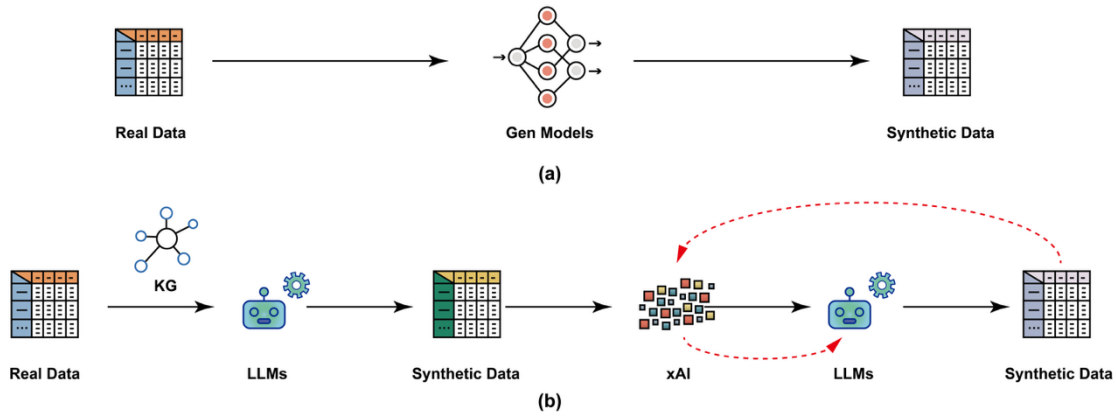

Figure 3: Comparison of generation paradigms: (a) conventional deep generative models; (b) our proposed KG- and xAI-guided LLM-based generation loop.

2. 関連研究

2.1 合成表形式データ生成

合成表形式データ生成は近年、プライバシー保護を目的としたデータ共有の必要性 [10][11]、堅牢なモデル開発 [9][14]、そしてデータが不足する環境におけるアルゴリズムのベンチマーキング需要の高まりから注目を集めている [18]。画像やテキストデータが空間的または時系列的な構造を持つのとは異なり、表形式データは異質な特徴を含み、混合データ型を扱い、さらに疎な依存関係やドメイン固有の依存関係を持つため、独自の課題を引き起こす [6][13][19]。

初期の研究では、ガウス・コピュラ [20][21]やベイジアンネットワーク [22][23]といった古典的な統計モデルが利用され、特徴量間の依存関係を因数分解あるいは直接モデル化することで結合分布を捉えようとした。しかし、これらのモデルは高次元の実データ環境では成立しにくく、特定の分布仮定や独立性仮定に依存している点に限界があった [9][14]。

その後、深層生成モデルの登場が大きな転換点となった。特に敵対的生成ネットワーク（Generative Adversarial Network: GAN）を基盤としたモデル、なかでもCTGAN [14]は表形式データに特化した条件付き生成メカニズムを導入し、条件付きサンプリングや戦略的な学習によりモード崩壊やクラス不均衡といった主要な課題に対応した。MedGAN [24]はこれを拡張し、マルチラベル医療記録の生成に適用したが、主に二値出力を対象としていた。近年では、TabDDPM [25]など拡散モデルを応用した手法が登場し、デノイジング・スコアマッチングを活用することで複雑なデータ多様体をより精緻に近似し、連続値データの生成で有効な結果を示している。

こうした進展にもかかわらず、多くのモデルは特徴量の周辺分布やペアごとの相関の保持に重点を置き、データセットに内在する高次の依存関係、論理的制約、ドメイン特有のルールを十分に考慮できていない [14][26]。またGANやVAEは訓練の不安定性や解釈可能性の欠如といった課題を抱え、金融や医療といった機微な領域での利用に制約が残っている [6][9]。

新しい方法として、大規模言語モデル（Large Language Model: LLM）を活用した合成データ生成が提案されている。GPT-3.5 や GPT-4 [27][28]などは、自然言語によるプロンプトを入力として構造化レコードを生成する能力を持ち、柔軟性と少量データでの利用可能性を備えている。実際に、LLM が JSON 形式のデータや表形式の行を訓練分布に沿って生成できることも報告されている [17]。しかし、これらのモデルはもともと表形式スキーマを理解するよう設計されていないため、生成結果には構造的一貫性を欠き、ハルシネーションや矛盾する特徴の組み合わせが生じやすい [29]。

さらに、統計的生成モデルと言語モデルベース生成モデルの両方に共通する問題として、ドメイン知識を十分に認識できないことが挙げられる。その結果、統計的には妥当であっても、意味的には不正確な出力が生成されることがある[30][31]。この制約を克服するため、近年では知識グラフのような構造化された事前知識を生成プロセスに統合する方向性が注目されており、本論文もこのアプローチをベースとする [16][32]。

2.2 データモデリングにおけるナレッジグラフ

ナレッジグラフ（Knowledge Graph: KG）は、セマンティック Web、オントロジー、企業データ統合など、多様な分野における構造化情報を表現する強力な抽象化手法として注目されている [16][32]。知識グラフは、現実世界のエンティティをノードとして表現し、それらの関係をラベル付きエッジとしてモデル化することで、事実の主張とドメイン論理の両方を表現できる。この表現は、多関係性、異質性、スキーマ柔軟性を備えた情報の符号化に適しており、表形式データの潜在的な構造と対応する。

構造化データモデリングの観点から、KG にはいくつかの利点がある。第一に、分類階層（例：職業はホワイトカラーの下位クラス）、依存関係（例：学歴が職種に影響を与える）、数値範囲（例：年齢が 0〜120 歳の範囲に収まる）といったドメイン制約を明示的にエンコードできる。これらは表形式データではしばしば暗黙的であるが、論理的一貫性のある合成レコードを生成するためには不可欠である。第二に、KG は意味的な推論や整合性チェックを可能にし、システムが矛盾を検出して修正できる。これは統計モデルには欠けている機能である[33][34]。

これまでの研究では、KG を活用したデータ拡張、エンティティリンク、関係学習などが検討されてきた。例えば、言語モデルに KG 埋め込みを統合して事実理解を強化する手法 [35]や、エンティティレベルの知識を組み込んで意味理解を向上させる言語モデル [36]が提案されている。しかし、データ合成の分野では KG の活用は依然としてほとんど未開拓である。ルールベースの生成やオントロジー制約に基づくサンプリングを試みた研究は存在するが [37]、様々な表形式データ領域への拡張性や汎用性に乏しいことが多い。

近年の TransE [38]、DistMult [39]、node2vec [40]などの知識グラフ埋め込み技術の発展は、記号的なグラフ構造を連続的なベクトル表現に変換する手段を提供し、下流のニューラルモデルが構造的知識を取り込むことを可能にしている。特に node2vec はランダムウォークを用いて局所的・全体的な文脈をバランスよく捉えられるため、表形式データをグラフに変換するシナリオに適している。こ

うした埋め込み技術により、生成プロンプトやモデル入力にグラフベースの意味情報を注入でき、生成サンプルの妥当性と解釈可能性を向上させられる [41]。

2.3 プロンプト設計と LLM ベースの生成

LLM の登場により、プロンプトを用いた条件付けによるゼロショットあるいは少ショットでの合成が可能となり、データ生成に新たな可能性が開かれた。従来のようにタスク特化の学習や微調整を必要とする生成モデルとは異なり、LLM はテキストによる指示や例を解釈することで、構造化データを含む多様なコンテンツを生成できる。この特性から、表形式データの合成に LLM を応用する試みが活発化している [17][42]。

プロンプトエンジニアリングとは、LLM に望ましい出力を導くために効果的な入力プロンプトを設計するプロセスを指す。当初は要約、翻訳、質問応答といった自然言語処理タスクに利用されていたが [28]、その後、表形式データの生成にも応用されている。例えば、構造化された JSON 形式のプロンプトを与えることで、LLM がスキーマ定義をある程度遵守するサンプルを生成できることが示されている [17]。また、列名・型・例示値を条件として与えることで、表形式のレコードを生成する指示調整型プロンプトの手法も提案されている [42]。

一方で、表形式データにおける LLM 生成には依然としていくつかの制約がある。大きな問題の一つは、LLM が構造化データや型付きデータに対する帰納的バイアスを持たない点である。そのため、生成結果には矛盾した値、欠落フィールド、スキーマの矛盾などの意味的な不整合が頻発する。例えば、患者記録の生成では年齢や性別と整合しない診断が含まれたり、請求記録ではビジネス制約に反する金額が出力される可能性がある。これは、LLM が本質的に自然言語コーパスで訓練されたシーケンス予測器であり、論理認識やスキーマ準拠を前提とした生成器ではないことに起因する [43]。

もう一つの課題は制御性の欠如である。LLM は表現力が高い反面、プロンプトの表現やフォーマット、例示のわずかな違いに敏感である。列間の依存関係を守る、稀なケースを再現する、といった特定の制約下でデータを生成するには、テンプレートを綿密に設計したり、複数回のサンプリングを行う必要があることが多い。さらに、プロンプトは通常ステートレスであり、長距離依存や複数レコード間の依存関係を強制するのが難しい [44]。

こうした背景から、近年は外部知識や構造を LLM の生成過程に組み込むハイブリッド手法が模索されている。オントロジーやデータ制約をプロンプトに埋め込み、事実に基づく出力を強化しようとする試みも報告されている [45]。しかし、それらの多くは経験則的あるいは特定ドメインに限られており、表形式データ合成における意味的一貫性を普遍的に制御できる手法には至っていない。

2.4 機械学習における説明可能性と SHAP

機械学習モデルが金融、医療、法律といった機微な分野に活用されるにつれ、モデルの解釈可能性は極めて重要な要件となっている。もはや正確な予測を提示するだけでは不十分であり、ステークホルダーはその予測の根拠について透明性と説明を求めるようになった [46][47]。こうした要請が説明可能な AI（Explainable AI: xAI）[48]の発展を促し、もともとブラックボックスであったモデルの意思決定プロセスを可視化・理解可能にする研究が進んでいる。

さまざまな xAI 手法の中でも、SHAP は事後的にモデルを説明するための代表的な枠組みとして広く利用されている [15]。SHAP は協力ゲーム理論に基づき、各入力特徴量が予測にどのように寄与したかをシャプレー値として定量化する。モデルに依存しない特性、理論的な基盤、そして局所的忠実性を備えており、モデル全体の挙動を一貫して解釈することが難しい場面でも有効である。

表形式データの領域でも、SHAP は臨床意思決定支援 [49]、信用スコアリング [50]、顧客離反予測 [51]といった多様なモデル解析に応用されている。個々の特徴量が予測に与える影響を可視化できるため、実務者はドメイン知識や規制要件に照らしてモデルの妥当性を検証できる。

合成データの文脈では、SHAP は予測精度だけでなく「意味的な整合性」を評価する視点を提供する。従来の評価指標（精度や F1 スコア、KL ダイバージェンスなど）はデータレベルやラベルレベルの類似性を測るにとどまる。これに対して SHAP を用いれば、合成データで訓練したモデルが実データで訓練したモデルと同じ特徴量の使い方をしているかを評価できる。これは極めて重要である。なぜなら、精度が高く見えても、モデルが異なる意思決定ロジックに依存していれば、実運用の際に偏りや安全性リスクを引き起こす可能性があるからである。近年の研究では、実データと合成データで学習したモデルの特徴重要度ベクトルを比較し、コサイン類似度、順位相関、ワッサーシュタイン距離などの指標で評価する方法が提案されている [52][53]。これらはモデル間の「意味的な隔たり」を捉え、統計的な分布の一致度だけでは見えない忠実度の差異をより精緻に測ることができる。

2.5 合成データの評価指標

合成表形式データの評価の重要性とは裏腹に、その評価は難しい。画像やテキストの生成タスクでは、人間による評価や事前学習済みモデルを用いた現実性の検証が可能である。しかし、表形式データには品質を判断する普遍的な基準が存在しないため、効果的な評価には統計的忠実度、下流タスクでの有用性、構造的一貫性、意味的整合性といった複数の次元を考慮する必要がある [54]。

一般的な方法のひとつは、実データと合成データの周辺分布を比較することである。KL ダイバージェンス、Jensen–Shannon ダイバージェンス（JSD）、Earth Mover's Distance（EMD）といった指標がよく使われ、各特徴量の分布の類似性を数値化できる [11]。平均や分散、エントロピーの差異を検出する点では有効だが、実際に重要となる多特徴量間の相互作用や高次依存関係を十分に捉えることは難しい。

この問題を補うため、ピアソン相関係数やスピアマン順位相関係数といった相関ベースの指標が用いられる。また、相互情報行列や部分相関構造などの高度な手法は、単純な周辺統計では見えない特徴間の細かな違いを捉えることができる [55]。しかし、これらも依然として統計的観点にとどまり、合成データが実際の予測環境でどのように機能するかを直接反映するものではない。

実務的には、タスクに特化した評価も広く利用されている。これは合成データでモデルを訓練し、実データでテストする、あるいはその逆を行う方法である。具体的には TSTR（Train on Synthetic, Test on Real）や TRTS（Train on Real, Test on Synthetic）といったプロトコルがある [9][14]。この設定では精度、F1 スコア、ROC-AUC、キャリブレーション曲線といった指標がよく報告される。ただし、これらはモデルの性能とデータの品質を区別しにくいため、予測精度は高くても意思決定のロジックが実データとは異なるケースを見落とす危険性がある。

この限界を克服する試みとして、近年は評価プロセスに説明可能性を取り入れる研究が進んでいる。SHAP 距離、特徴重要度の順位相関、帰属分布間のワッサーシュタイン距離といった指標は、新たに意味的一貫性を測る試みが行われている [15][52]。これらは実データと合成データで訓練したモデルが学んだ推論過程を比較できる。例えば、予測に大きな影響を与える特徴量の上位 3 つがモデル間で異なる場合、精度スコアが似ていても意味的なギャップがあることを示唆する。また、SDMetrics[3]のような包括的フレームワークは、単変量・多変量の統計評価、予測性能の評価、人間による解釈を支援する可視化機能などを統合している。しかし、こうしたツールでもドメインロジックとの整合性や制約の妥当性を完全に保証することはできない。これらはまさに本研究が取り組むべき課題である。

まとめると、合成データの評価指標は多く存在するものの、生成品質の改善に直接つながる指針を提供するものは少ない。本研究で提案するフレームワークは、統計的および意味的な整合性を評価するだけでなく、その情報を生成ループにフィードバックし、合成データの質を反復的に向上させることを目的としている。

[3] https://docs.sdv.dev/sdmetrics

3. 提案手法

3.1 本研究の問題定義

本研究の目的は、統計的および意味的の両側面において、実世界のデータの高忠実度な代替となる合成表形式データを生成することである。具体的には、実データセットD_{real}に対し、以下の条件を満たす合成データセットD_{syn}を構築することを目指す。

1. D_{syn}の統計的分布がD_{real}の分布に近似していること
2. D_{syn}で訓練された機械学習モデルの決定挙動が、D_{real}で訓練されたモデルの挙動と密接に一致すること

実データセットを$D_{\mathrm{real}} = \{x_i\}_{i=1}^{N}$とする。各サンプル$x_i$は、$d_c$個の連続特徴量と$d_{\mathrm{cat}}$個のカテゴリ特徴量から構成される。これを模倣する合成データセットを$D_{\mathrm{syn}} = \{x_i'\}_{i=1}^{N'}$と定義する。

f_{real}およびf_{syn}は、それぞれD_{real}およびD_{syn}で独立に学習した機械学習モデルとする。各インスタンスxに対して、$\phi_{\mathrm{real}}(x)$および$\phi_{\mathrm{syn}}(x)$を SHAP による特徴属性ベクトルとし、それぞれのモデルから算出する。各次元$\phi_j(x)$は、特徴量jがモデル出力に与える寄与度を表す。実データと合成データで学習したモデル間の属性ベクトルの違いを定量化するため、次の指標を定義する。

$$\mathrm{SemanticGap} = E_{x\sim\mathcal{D}}\left[D\left(\phi_{\mathrm{real}}(x), \phi_{\mathrm{syn}}(x)\right)\right]$$

- $\mathcal{D}$：データ分布（例：D_{real}）
- E：期待値（サンプル平均）
- $\phi_{\mathrm{real}}(x)$：実データモデルの SHAP ベクトル
- $\phi_{\mathrm{syn}}(x)$：合成データモデルの SHAP ベクトル
- $D(\cdot,\cdot)$：2 つのベクトル間の距離関数

本研究では、距離関数Dとしてコサイン距離を用いる。結果は 0〜2 の範囲をとり、0 は完全一致を示す。

$$D_{\cos}\left(\phi_{\mathrm{real}}, \phi_{\mathrm{syn}}\right) = 1 - \frac{\phi_{\mathrm{real}} \cdot \phi_{\mathrm{syn}}}{|\phi_{\mathrm{real}}| \cdot |\phi_{\mathrm{syn}}|}$$

したがって、データセット全体における意味的ギャップは次のように表される。

$$\mathrm{SemanticGap}_{\cos} = \frac{1}{N}\sum_{i=1}^{N}\left(1 - \frac{\phi_{\mathrm{real}}^{(i)} \cdot \phi_{\mathrm{syn}}^{(i)}}{|\phi_{\mathrm{real}}^{(i)}| \cdot |\phi_{\mathrm{syn}}^{(i)}|}\right)$$

3.2 全体フレームワークの概要

KGSynX フレームワークは、(i) 知識グラフによる構造モデリング、(ii) LLM を用いたプロンプトベースの生成、(iii) 属性に基づく説明可能なフィードバックという 3 つの主要コンポーネントを統合し、合成表形式データを生成するよう設計されている。図 4 は、提案する KGSynX の全体構成を示している。本フレームワークは主に4つの段階からなり、(1) 構造化データを知識グラフへ変換、(2) グラフ埋め込みと LLM による初期合成、(3) SHAP 解析とプロンプトフィードバックループ、(4) 最終的な合成データ生成、という流れとなる。

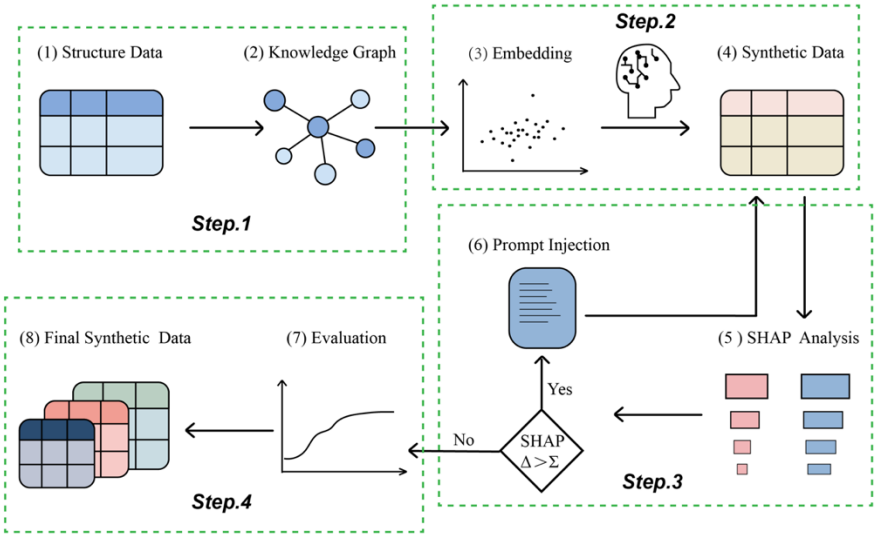

Figure 4: Overview of the KGSynX framework.

3.3 知識グラフ構築

表データを平坦にせず、各行をエンティティ、各「属性–値」ペアをノードとして知識グラフ化する。本研究では、型付き・有向エッジによって① has_attribute（行→属性値）、② is_a（階層）、③ co_occurs（共起依存）、④ 数値制約を表現した。連続値は区間でアノテーションし、カテゴリは直接ノード化する。これにより、小規模データでも稀な組合せやドメイン規則といった構造知識を抽出でき、医療・金融など多様な領域へ柔軟に拡張・保守可能となる。

具体的なプロセスを以下に示す。まず、テーブルデータセット$D_{\text{real}} = \{x_i\}_{i=1}^{N}$の各レコード$x_i \in \mathbb{R}^{d_c} \times C^{d_{\text{cat}}}$は、一連の連続特徴量とカテゴリ属性で構成される。ここで、連続特徴ベクトル$\boldsymbol{x}_i^{(c)} \in \mathbb{R}^{d_c}$は実数値の測定値（例：年齢、収入）で構成され、一方、カテゴリ属性ベクトル$\boldsymbol{x}_i^{(\text{cat})} \in C^{d_{\text{cat}}}$は有限集合から抽出された離散ラベル（例：性別、国）で構成される。続いて、スキーマ$S = \{A_1, A_2, \cdots, A_d\}$を抽出する。各属性$A_j$は可能な値を表す領域と関連付けられる。各属性と値のペ

ア(A_j, a_k)は意味の単位として扱われ、a_kは属性A_jの領域$\mathbb{D}_j$におけるk番目の値を示す。カテゴリ特徴は直接ノードにマッピングされ、連続特徴は範囲メタデータ（例：“age: 30-39”, “$income > \$100k$”）を用いてアノテーションされる。

表 1 は本研究の実験でも用いる UCI Heart Disease データから抽出した 3 つのサンプルであり、図 5 は上記の手法によって得られた知識グラフである。

Table 1: UCI Heart Disease データのサンプル

Feature	Sample 0	Sample 1	Sample 2
age	63	67	67
sex	1	1	1
cp	1	4	4
trestbps	145	160	120
chol	233	286	229
fbs	1	0	0
restecg	2	2	2
thalach	150	108	129
exang	0	1	1
oldpeak	2.3	1.5	2.6
...	...	...	...
thal	6	3	7
target	0	1	1
sex_label	male	male	male
cp_label	typical_angina	asymptomatic	asymptomatic
fbs_label	true	false	false
restecg_label	left_vent_hyper	left_vent_hyper	left_vent_hyper
exang_label	no	yes	yes
slope_label	downsloping	flat	flat
thal_label	fixed_defect	normal	reversible_defect
ca_label	0_vessels	3_vessels	2_vessels

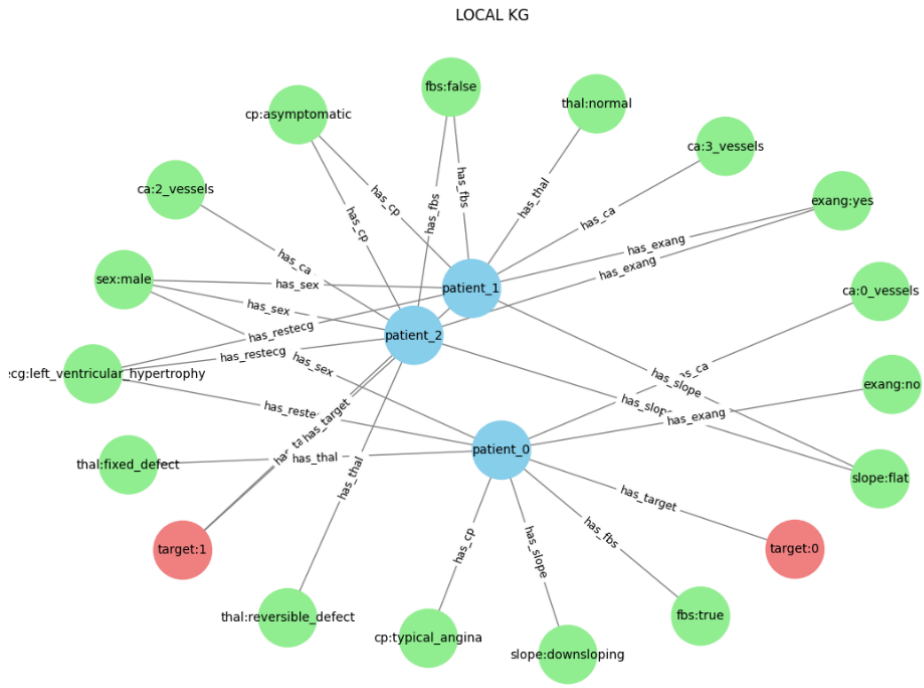

Figure 5: A local knowledge graph derived from three UCI Heart Disease samples.

3.4 埋め込みと LLM ベースのデータ生成

知識グラフを node2vec 等のグラフ埋め込み手法で埋め込み化し、構造・関係情報を保持したベクトルをプロンプト設計に用いる [40][56]。このプロンプトによって生成された表形式データは、平文のみの指示よりスキーマ整合性・ドメイン論理に沿った出力になりやすいことが知られている [17][57]。さらに埋め込みは特徴–値ペアのクラスタリング／類似度計算を可能にし、「契約=長期」と「勤続年数=高」など関連概念を近くに配置して、一貫した文脈要素の選択を行えるようになる。

構築したプロンプトを事前学習 LLM に入力し、知識グラフにて構造化された制約を含めることで、表形式レコードの生成におけるスキーマ整合性とドメイン妥当性を高める。先行研究[17][58]では、プロンプトベースの手法が構造化生成タスクに有効であることをと示す一方、手作業テンプレートやテキスト中心のヒューリスティクスに依存していた。本手法はグラフ構造と埋め込みを統合し、記号表現とニューラル生成の橋渡しするプロセスを有する。厳密な論理関係が重要となる医療や金融分野では、ハルシネーションや矛盾が合成データの利用可能性を低下させてしまうため[59]、本アプローチは非常に有効であると考える。以下が本研究で用いたプロンプトの例である。

Initial Prompt Template

```
prompt = (
      "You are a data generation assistant.¥n"
      "Generate a new synthetic patient record based on the "
      "following structural features.¥n"
      "Only include the following fields (no extra fields):¥n"
      f"{required_fields}¥n¥n"
      "Structural features:¥n"
      + "¥n".join(structured_features)
      + "¥n¥nReturn a single JSON object only.¥n"
)
```

3.5 意味的フィードバックループ

もっとも、知識グラフと構造化したプロンプトだけでは、実データで訓練したモデルと一致した結果を得られる保証はない。そこで、属性情報に基づく改善ループ（意味的フィードバックと呼ぶ）を導入する。具体的には、実データと合成データでそれぞれ訓練した同型のモデルに対して SHAP を適用し、特徴重要度ベクトルを算出し、比較する。本研究では、両者の差異を意味的ギャップと呼ぶ。ギャップが大きいとは、表面的な精度が近く見えても、予測に寄与する特徴の重み付けが異なることを意味する。このギャップを手掛かりに、過小評価されている、あるいは誤って表現されている特徴を特定し、次ラウンドの生成でプロンプトを修正する。閾値以下であれば合成データは意味的に一貫していると判断され、最終出力のステップへと進む。

図 6 は意味的整合性を反復して改善するプロセスの概念図を示す。また、アルゴリズム 1 は意味的ギャップを得るプロセスの擬似コードである。意味的ギャップが低ければ、合成データが類似した予測ロジックを支持していることを意味する。一方、ギャップが高い場合、精度が類似して見えても、合成データで訓練されたモデルが異なるパターンを学習していることを示す。なお、反復ループは、以下のいずれかの条件を満たした時点で終了する。

- 意味的ギャップの改善度が閾値$\Delta < \Sigma$を 2 ラウンド連続で下回った場合
- 反復回数の上限Tに達した場合
- 下流モデルの精度が低下し、SHAP 損失への過学習を示唆する場合

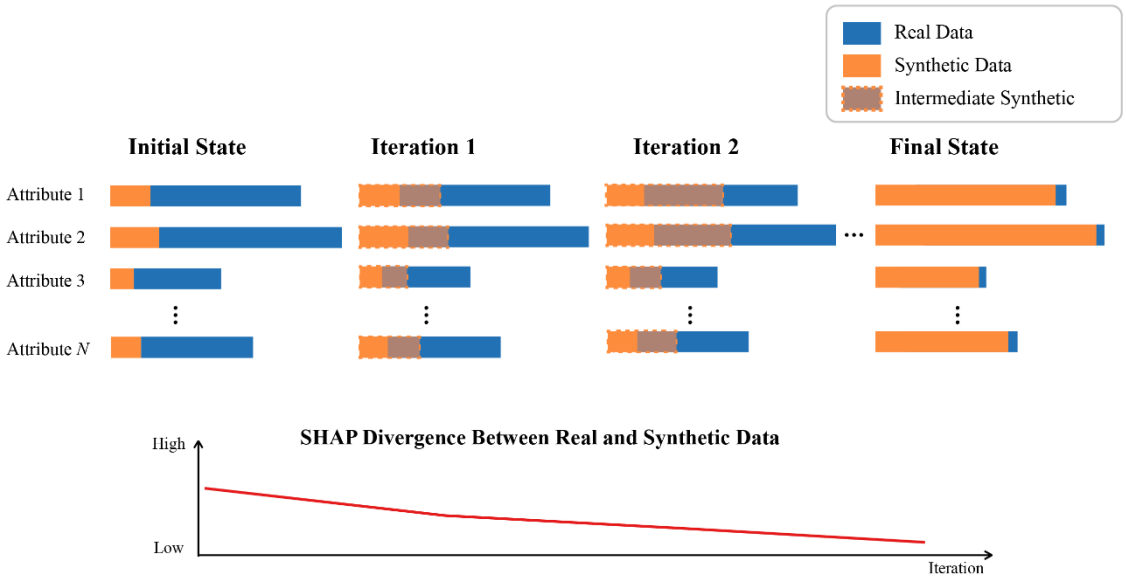

Figure 6: SHAP alignment process during prompt refinement.

Algorithm 1 Compute Semantic Gap

Require: real dataset D, models model_real, model_syn
Ensure: semantic gap value

$sum \leftarrow 0$
$N \leftarrow |D|$
for all $x \in D$ **do**
  $\phi_{\text{real}} \leftarrow \text{SHAP}(\text{model\_real}, x)$
  $\phi_{\text{syn}} \leftarrow \text{SHAP}(\text{model\_syn}, x)$
  $gap \leftarrow 1 - \dfrac{\phi_{\text{real}} \cdot \phi_{\text{syn}}}{\|\phi_{\text{real}}\| \, \|\phi_{\text{syn}}\|}$
  $sum \leftarrow sum + gap$
end for
RETURN sum/N

3.6 評価と最終出力結果の取得

精緻化された合成データは、統計的・構造的・意味論的観点から最終評価を受ける。すべての評価基準を満たした場合、そのデータセットは機械学習の学習、ベンチマーク、データ共有といった下流タスクに利用可能な高品質な合成表形式データと見なされ、最終出力結果として出力される。

以下のアルゴリズム 2 が KGSynX の全体のプロセスを表す疑似コードである。

Algorithm 2 KGSynX: Knowledge Graph-Guided and SHAP-Refined Tabular Data Generation

Require: Real dataset D_{real}, LLM $\mathcal{M}$, SHAP threshold Σ, max rounds T
Ensure: Synthetic dataset D_{syn}

Build Knowledge Graph G **from** D_{real}
for each sample $x_i \in D_{\text{real}}$ **do**
    Add entity node v_i, link to attribute nodes via typed edges
end for
Train Node2Vec on G to obtain embeddings $\mathbf{z}_i$ for each entity
for $i = 1$ to T **do**
    Generate prompt $\mathcal{P}_i$ from $\mathbf{z}_i$
    Generate synthetic sample $x_i' \leftarrow \mathcal{M}(\mathcal{P}_i)$
    $D_{\text{syn}}^{(i)} \leftarrow D_{\text{syn}}^{(i)} \cup \{x_i'\}$
    Train classifier $\mathcal{C}_{\text{real}}$ // trained on real data D_{real}
    Train classifier $\mathcal{C}_{\text{syn}}$ // trained on synthetic data D_{syn}
    Compute SHAP vectors $\phi_{\text{real}}, \phi_{\text{syn}}$
    if $\|\phi_{\text{real}} - \phi_{\text{syn}}\|_{cos} < \Sigma$ **then**
        Break
    else
        Update prompt focus based on top-k SHAP gaps
    end if
end for
return D_{syn}

4. 実験

4.1 実験設定

KGSynX の有用性を評価するため、医療、企業取引、電気通信分野から 3 種類の構造化表形式データセットを用いて実験を行った。これらは、実社会における合成表形式データの代表的な分野であり、プライバシーや意味的整合性の確保が主要な課題となっている [6][14][60]。

KGSynX では、各ラウンドで実データの 1 件のレコードに対して、正確に 1 件の合成記録を生成する。生成ループは以下のいずれかの条件で終了する。

(a) SHAP の意味的整合性ギャップが閾値Σ（デフォルト 0.1）を下回った場合

(b) 最大ラウンド数T（デフォルト 5）に達した場合

実際の運用では、データ忠実度と API コストのバランスを考慮し、3〜4 ラウンド以内に収束した。すべての合成データは ChatGPT-4o によって生成され、データセットあたり平均 20 USD のコストがかかった。これは、企業や研究機関における利用において十分実用的と考えられる。

4.2 データセット

UCI Heart Disease データ：このデータセットはクリーブランド・クリニック財団による 303 名の患者の生体医学記録である [61]。各レコードには年齢、コレステロール値、安静時血圧、心電図結果などの属性を含んでいる。二値ターゲットは患者が心臓疾患を有するか否かを示す。小規模かつ不均衡な性質から、本データセットは医療的に敏感でリソースが限られたシナリオにおける生成手法のテストに最適である。

企業向け請求書利用状況データ：この独自データセットは、株式会社インフォマートの提供している日本の企業向け SaaS 請求書プラットフォームを利用する 17,015 社の行動記録である。月次請求書発行数（InvoiceCount_1Month）、6 か月間の総請求書利用数（InvoiceCount_6MonthTotal）、総ユーザー数、ID、パートナータイプ、無料および有料ユーザー数が含まれている。二値ターゲット変数 InvoiceIssued は、企業が有料請求書を発行したかどうかを示す。このデータセットは現実世界の企業行動を反映しており、構造化データ合成における意味の一貫性を評価するのに適している。

通信事業者の顧客離反データ：この公開データセットは、通信サービスプロバイダーの 7,000 人以上の顧客に関する契約および請求詳細である[4]。カテゴリ変数（例：Contract, InternetService）と数値変数（例：MonthlyCharges）の両方を含む。予測課題は顧客の解約判定である。特徴量の多様性と規模から、混合データ型下での合成生成の頑健性を評価する優れたベンチマークとなる。

4.3 ベースライン手法

従来型ベースラインとして、表形式データ合成向けに最適化された 2 つの広く採用されている生成的敵対ネットワーク（GAN）である CTGAN [14]と MedGAN [24][62]を用いる。実験では公式オープンソース実装を採用し、推奨設定に従ってパラメータを調整した。

また、最近提案された表形式データに特化した拡散ベースの生成モデルである TabDDPM [25]も採用した。GAN が敵対的学習に依存し不安定性やモード崩壊に陥りやすいのとは異なり、拡散モデルはノイズを反復的に構造化されたサンプルへと精緻化するノイズ除去処理を通じてデータ分布を学習する。

モデルベース学習アプローチとは対照的に、我々はゼロショットまたは少ショットの表形式データ生成に LLM を用いている。実験における全ての LLM ベース手法は、OpenAI API 経由でアクセスする OpenAI の ChatGPT-4o モデルをベースとする。これにより、LLM ベースのバリエーション間で言語生成品質とアーキテクチャ挙動の一貫性が保証される。

実験では 2 つの LLM ベースラインとして、LLM-only と LLM+KG を用いる。LLM-only では言語モデルの純粋な生成能力を評価し、LLM+KG では、知識グラフに基づくプロンプトで生成プロセスを強化したものという位置づけである。

4.4 評価プロトコルと評価指標

TSTR プロトコルに従って、分類器は合成データで訓練され、実データセットから除外したテストセットで評価される。このプロトコルは、合成データが実データと同等に下流タスクをサポートできるかを直接評価する。80：20 分割を使用し、異なる乱数シードで各実験を 5 回繰り返した。報告結果は全試行の平均値である。

合成データの品質を予測性能と意味的整合性の 2 観点から評価する。前者は精度、精緻度、再現率、F1 スコア、AUC といった標準分類指標を採用し、TSTR プロトコル下で算出する。この設定では分類器を生成データで訓練し、除外され

[4] https://www.kaggle.com/datasets/blastchar/telco-customer-churn

た実データでテストするため、合成記録が実データと同等の正確な下流推論を可能にするかを評価できる。

表面的な性能評価に加え、合成データで訓練されたモデルが実データで訓練されたモデルと同様の決定ロジックを学習しているかどうかの評価を重視する。この目的のため、SHAP を用いてグローバルな特徴量を抽出する。

実データ学習モデルと合成データ学習モデル間の意味論的一貫性を定量化するため、SHAP 帰属ベクトル間のコサイン距離に基づくセマンティックギャップスコアを算出する。この指標は、同一予測課題に対する両モデルの推論類似性を算出する。KL ダイバージェンス [63]やワッサーシュタイン距離 [64]などの従来の分布間距離とは異なり、セマンティックギャップは統計的類似性だけでなく特徴量使用における機能的整合性も反映するため、解釈可能性と信頼性が重要な領域において有用である。

4.5 分類器

合成データの下流タスクにおける有用性を評価するため、生成サンプル上で機械学習分類器を訓練し、TSTR プロトコルを用いて保持された実データでテストする。全データセットに単一分類器を適用する代わりに、各分野の特性とモデリング慣習に適合したモデルを選択し、現実的で意味のある評価を行う。

UCI Heart Disease データセットでは、解釈可能性、ノイズに対する頑健性、過学習への耐性を兼ね備えた小規模生体医学データに適したアンサンブル手法であるランダムフォレスト分類器 [65]を採用した。ランダムフォレストはカテゴリ変数と連続変数を含む異種特徴量を効果的に処理し、SHAP ベースの説明モデルを自然にサポートするため、意味的整合性の評価に適している。

企業向け請求書利用状況データでは、クラス不均衡のある表形式ビジネスデータの処理において優れた性能を発揮する XGBoost [66]を選択した。このデータセットにおける二値分類課題（企業が有料請求書を発行するか否か）は本質的に偏っている。XGBoost はモデルの複雑度と正例重み付け scale_pos_weight パラメータ経由）を微調整可能であり、不均衡な設定で効果を発揮する。その木構造は、アクティブユーザー数や請求頻度といった利用特徴量間の複雑な閾値相互作用も捉えられる。

通信事業者の顧客離反データでは、中規模構造化データでの性能に最適化された勾配ブースティング決定木フレームワークである LightGBM [67]を採用する。本データセットにはサービス属性、請求情報、契約タイプが混在している。このようなデータタイプにおいて LightGBM は高い効率性と精度を実現しうる。また、顧客離反予測タスクにおける事実上の標準となっている。

表 2 に実験で用いた諸手法のパラメータをまとめる。

Table 2: Summary of Experimental Parameters

Category	**Parameter**	**Value (Default)**
Random Forest	n_estimators	100
	criterion	gini
	max_depth	None
	min_samples_split	2
	min_samples_leaf	1
	bootstrap	True
XGBoost	n_estimators	100
	max_depth	6
	learning_rate	0.3
	subsample	1.0
	colsample_bytree	1.0
	reg_alpha / reg_lambda	0 / 1
LightGBM	num_leaves	31
	max_depth	-1
	learning_rate	0.1
	n_estimators	100
	feature_fraction	1.0
	bagging_fraction / freq	1.0 / 0
	min_data_in_leaf	20
LLM	Temperature	0.8
	Top-p	0.95
	Max tokens	300
Refinement Loop	T (iterations)	5
	k (top features)	10
	Σ (threshold)	0.1

5. 結果と考察

5.1 UCI Heart Disease データにおける性能

UCI Heart Disease データにランダムフォレスト分類器で評価した結果、KGSynX は全ての合成データ手法の中で最も高い性能を示し、F1 スコア 0.750 を達成した（表 3）。これは実データで得られる上限値（F1: 0.826）との差を大きく縮める結果である。特に注目すべき点は、KGSynX が CTGAN およびプロンプトのみを用いた LLM ベースラインの双方を上回ったことである。AUC の指標でも、CTGAN の 0.746 から KGSynX の 0.827 へと大きく向上した。これは、SHAP を利用したフィードバックによる改良が、実データに見られる決定境界に沿った学習能力を合成データ生成モデルに付与したためと考えられる。

Table 3: Heart Disease Dataset (RF)

Method	**Acc**	**Precision**	**Recall**	**F1**	**AUC**
Real	0.867	0.864	0.792	0.826	0.929
MedGAN	0.664	0.298	0.541	0.384	0.527
CTGAN	0.667	0.643	0.375	0.474	0.746
TabDDPM	0.541	0.293	0.557	0.380	0.498
LLM	0.350	0.297	0.458	0.361	0.278
LLM+KG	0.600	0.500	0.833	0.625	0.741
Ours	**0.767**	**0.656**	**0.875**	**0.750**	**0.827**

5.2 企業向け請求書利用状況データにおける性能

企業向け請求書データに XGBoost による分類を行った結果、KGSynX は精度（0.900）、F1 スコア（0.904）、精確度（0.870）、再現率（0.940）の 4 つの主要指標すべてにおいて最も高い性能を示した（表 4）。LLM+KG ベースラインが AUC で 0.943 と最良の結果を達成した一方、KGSynX も AUC が 0.942 とほぼ同等の高い性能を示した。

これらの結果は、企業固有の構造的文脈を知識グラフとして組み込み、さらに SHAP に基づくフィードバック精緻化を加えることで、実ビジネス環境における合成データの有用性が大幅に向上することを裏付けている。また、CTGAN との F1 スコアにおける 23.0%という大きな性能差は、意思決定ロジックがドメイン依存となる制約付き機密データにおいて、グラフ構造による事前知識と意味的フィードバックが持つ優位性を有していると考えられる。

Table 4: Enterprise Invoice Dataset (XGBoost)

Method	Acc	Precision	Recall	F1	AUC
Real	0.867	0.778	0.700	0.826	0.929
MedGAN	0.725	0.726	0.674	0.724	0.818
CTGAN	0.655	0.642	0.700	0.670	0.628
TabDDPM	0.425	0.371	0.637	0.357	0.544
LLM	0.765	0.762	0.770	0.766	0.838
LLM+KG	0.865	0.848	0.890	0.868	**0.943**
Ours	**0.900**	**0.870**	**0.940**	**0.904**	0.942

5.3 通信事業者顧客離反データセットにおける性能

顧客離反データセットにおける LightGBM を用いた分類実験では、KGSynX が他の手法を上回る性能を示した（表 5）。一方、LLM のみのベースラインは最高の再現率（0.929）を記録したが、精度が大幅に低下したため全体としては不安定な性能となり、F1 スコアは低かった。これは、言語モデルが局所的な属性パターンを捉えることは可能であっても、外部からのガイダンスがなければ全体的な一貫性を維持するのが難しいことを示している。一方で、KGSynX は精度と再現率のバランスを保ち、意味的制約と SHAP によるフィードバック精緻化を組み込むことで、数値・カテゴリを含む混合型のデータセットにおいて安定性と汎化性能の双方を向上させられている。

Table 5: Telco Customer Churn Dataset (LightGBM)

Method	Acc	Precision	Recall	F1	AUC
Real	0.833	0.689	0.738	0.713	0.867
MedGAN	0.730	0.501	0.530	0.515	0.294
CTGAN	0.726	0.531	0.241	0.332	0.557
TabDDPM	0.721	0.518	0.610	0.603	0.772
LLM	0.626	0.425	**0.929**	0.584	0.810
LLM+KG	0.760	0.584	0.206	0.326	0.824
Ours	**0.776**	**0.662**	0.901	**0.611**	**0.853**

5.4 アブレーション研究

KGSynX フレームワークの各コアコンポーネントの貢献度を明らかにするため、アブレーション研究を行った。個別のモデルを比較するのではなく、フレームワークを以下の 3 段階に分解して性能を評価した : (1) 構造やフィードバックを持たないプロンプトベース生成（LLM のみ）、(2) 知識グラフによるプロンプト構築（LLM+KG）、(3) SHAP を用いたフィードバックによる生成（KGSynX）。

5.4.1 UCI Heart Disease データ

LLM のみから LLM+KG へ移行することで、F1 スコアは 0.361 から 0.625 に、再現率は 0.458 から 0.833 へと大幅に改善した。これは、構造化されたプロンプトが医学的に意味のある特徴の組み合わせを適切に捉えていることを示す。さらに SHAP ベースのフィードバックを導入すると、F1 スコアは 0.750、再現率は 0.875 に上昇し、実データの性能に近づいた。

5.4.2 企業向け請求書利用状況データ

UCI Heart Disease データと同様の傾向が見られ、LLM のみでは F1 スコア 0.766、KG 条件付けで 0.868、フィードバック導入で 0.904 へと順に改善した。特に SHAP フィードバックによる調整後は精度と再現率のバランスが顕著に向上した。また、本データセットの課題は主にその構造的特徴、すなわちカテゴリの疎さ、特徴量の歪んだ分布にある。これらのビジネス固有のロジックが合成生成を特に困難にした。それにもかかわらず、KGSynX は分類性能と一貫性において大幅な改善を達成した。

5.4.3 通信事業者の顧客離反データ

LLM のみのモデルは非常に高い再現率を達成したが、精度が低く、F1 スコアも 0.584 にとどまり汎化性能に課題があった。KG を導入した生成は構造的には優れていたが、再現率が著しく低下した。これに対し、両コンポーネントを統合した KGSynX では、全ての指標においてバランスの取れた結果（F1: 0.611, AUC: 0.853）が得られた。

6. 合成データの品質分析

6.1 UCI Heart Disease データ

6.1.1 周辺分布と KL ダイバージェンス

まず、age、trestbps、chol、thalach といった主要な数値特徴量の周辺分布を分析した。実データと合成データの分布を比較した結果、合成データは実データの中央傾向とは概ね一致しているものの、中央部分のピークが強くなり、尾部の分散が減少する傾向が確認された（図 7）。

続いて整合性を定量的に評価するため、主要な変数の KL ダイバージェンスを算出した（表 6）。その結果、一部の特徴量では許容範囲内の整合性が確認されたが、特に血圧（trestbps）、ST 低下（oldpeak）、年齢分布において顕著な乖離が認められた。

Table 6: KL divergence (UCI Heart Disease dataset)

Feature	KL Divergence
age	5.3840
trestbps	6.0128
chol	0.9035
thalach	0.4967
oldpeak	4.6848

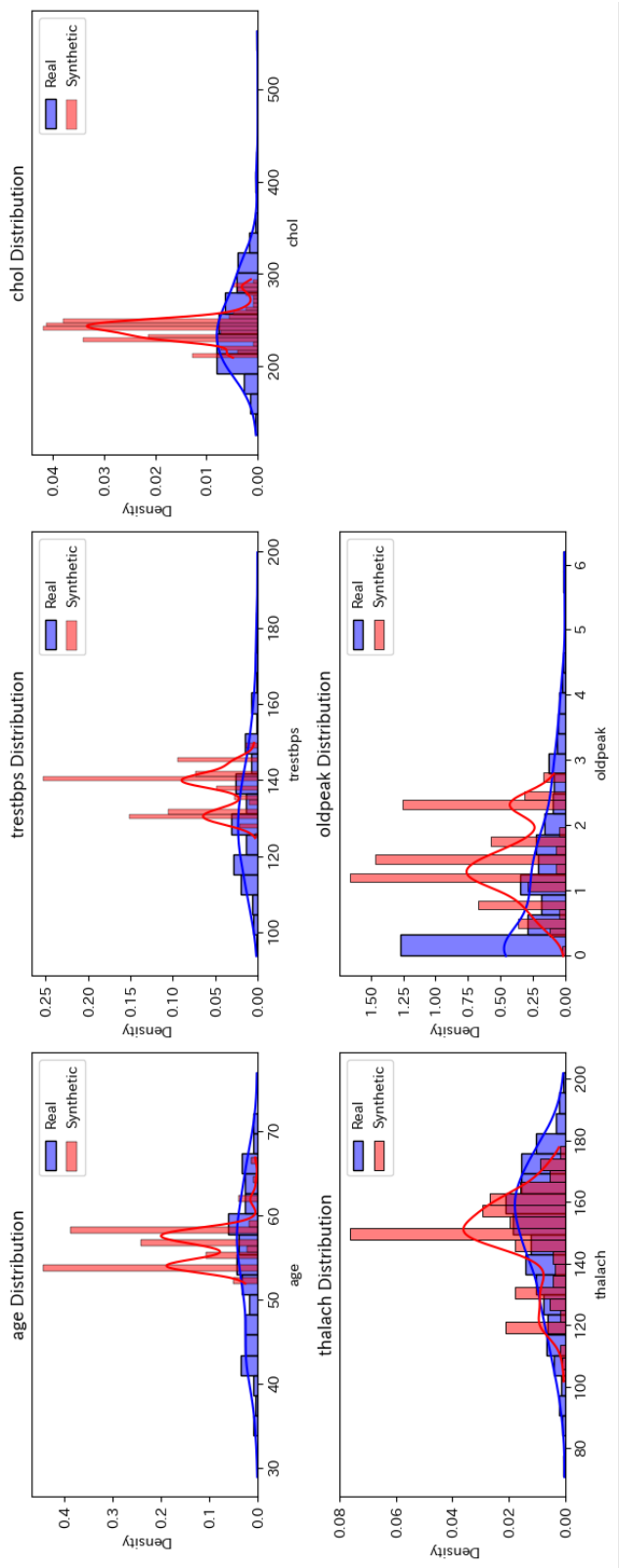

Figure 7: Density comparison (UCI Heart Disease).

6.1.2 データ構造の PCA 可視化

実データと合成したデータの大域的な構造を比較するため、両者を主成分分析（PCA） [68]によって二次元空間に可視化した。分析の結果、実データは投影平面上で複数方向に広がり、空間的な分布がより多様であることが確認された。一方で、KGSynX が生成した合成データは投影空間の重心付近に集中する傾向が強く見られた（図 8）。

これは、KGSynX による生成が主要な統計的特徴の保持には有効である一方、周辺部や低密度領域の表現が不十分であることを示唆している。こうした挙動は、生成モデルにおけるモード崩壊やモード切り捨てに関する先行研究の知見と一致しており [69][70]、とりわけプロンプトが一貫性と妥当性を優先して設計されている場合に顕著に現れることが知られている。特に、LLM を用いたプロンプトベースの生成は高頻度かつ整ったパターンに偏りやすく、稀な事例や特徴の組み合わせを十分に再現できないことが多い [17]。

続いて、この合成データの傾向を定量的に調べるため、最初の 2 つの主成分が説明する分散比率を比較した（表 7）。実データでは、PC1 が 35.94%、PC2 が 21.94%（合計 57.88%）であったのに対し、合成データでは PC1 が 64.84%、PC2 が 13.89%（合計 78.73%）となった。合成データにおける PC1 の分散比率の顕著な増加は、実データのような多方向へ広がらず、単一の軸方向に強く集中する傾向を定量的に裏付けている。この変化は、合成データ生成において完全な共分散構造を保持する仕組みの必要性を示していると言える。

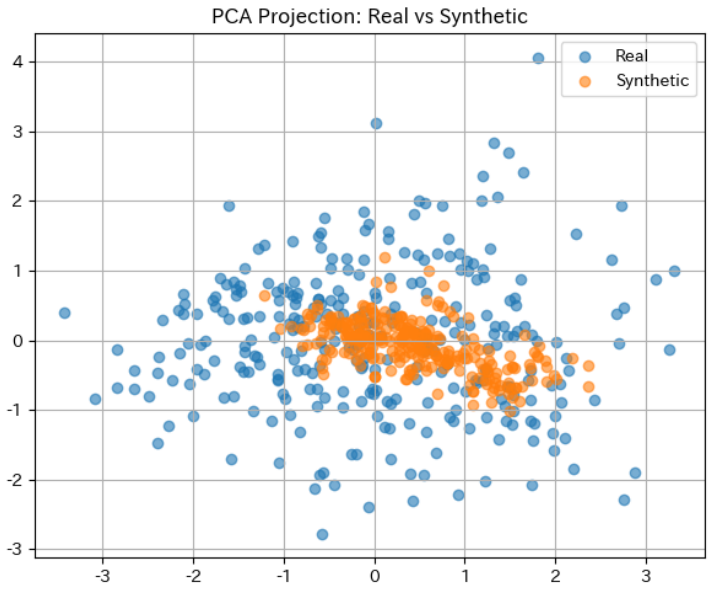

Figure 8: PCA projection (UCI Heart Disease).

Table 7: Explained variance ratio of the first two principal components (UCI)

Dataset	PC1 (%)	PC2 (%)	PC1+PC2 (%)
Real data	35.94	21.94	57.88
Synthetic data	64.84	13.89	78.73

6.1.3 SHAP による属性アライメント

ここでは実データと合成データそれぞれで訓練したランダムフォレスト分類器に対して SHAP 値を計算し、影響力の大きい上位 10 件の特徴量を比較した（図 9）。その結果、thal、thalach、cp などの特徴量は両モデルで一貫して主要な予測因子として現れ、高い整合性を示していた。一方で、合成データで訓練したモデルでは、oldpeak や trestbps といった特徴量で SHAP 値が上昇しており、これはプロンプトによるバイアス、あるいは生成データにおいて自然な相関を再現できていないことを反映していると考えられる。

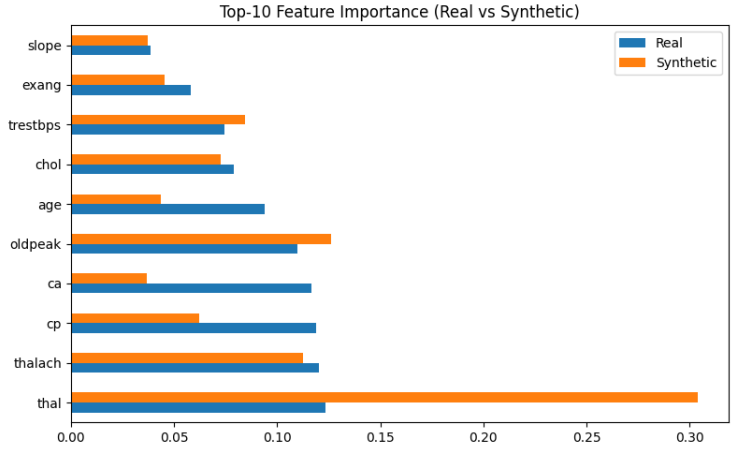

Figure 9: SHAP feature attribution comparison (UCI Heart Disease).

6.1.4 SHAP に基づく特徴量帰属の一貫性分析

合成データが統計的特性だけでなく、下流分類器に反映される意味的な決定ロジックも保持しているかを検証するため、実データと合成データで学習したモデルの SHAP 値を比較した。本分析では、グローバルな帰属の一貫性、局所的な相互作用の忠実度、そして生成プロセスに起因する潜在的な偏りに注目した。

実データに基づく SHAP のプロットから、cp、ca、thal といったユーザー関連の特徴量がモデルの予測に強く寄与していることが確認できる（図 10）。特に重要なのは、SHAP 値の方向性がドメイン知識と整合している点である。例えば、

oldpeak、ca、thal の値が高いほどリスク予測も高くなる傾向が見られる。さらに依存性プロットの分析では、age と chol の間に非線形な相互作用が存在しており、異なるサブ集団ごとに SHAP 値が変化している様子が見られた（図 11）。

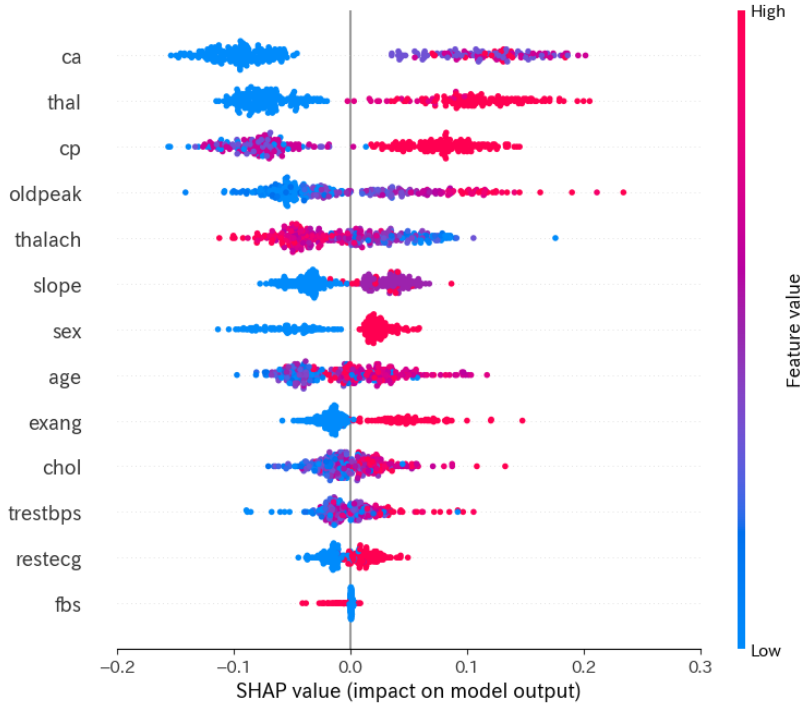

Figure 10: SHAP summary plot for real UCI Heart Disease data.

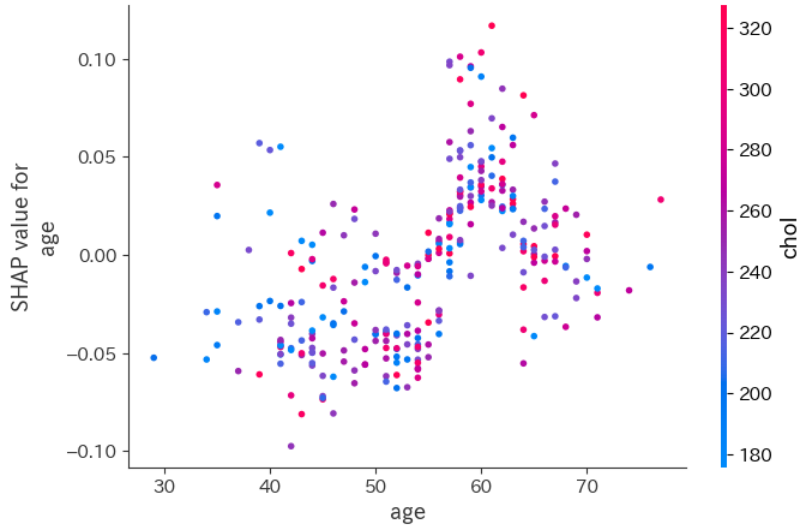

Figure 11: SHAP dependence plot for real data: effect of age (colored by cholesterol).

図 12 は合成データの SHAP 値プロットである。図 10 の実データのものと比較し、合成データはこれらのパターンの多くを再現している様子が見られる。特徴量の帰属は概ね一貫性を保ち、合成モデルにおける上位特徴量は実データモデルのものと一致しており、意思決定ロジックの全体構造がよく保存されていることを示している。合成データの SHAP 依存性プロットも、コレステロール値を介した年齢と予測リスクの正の相関関係を再現している（図 13）。一方で、分布の正則化により範囲が圧縮されている。

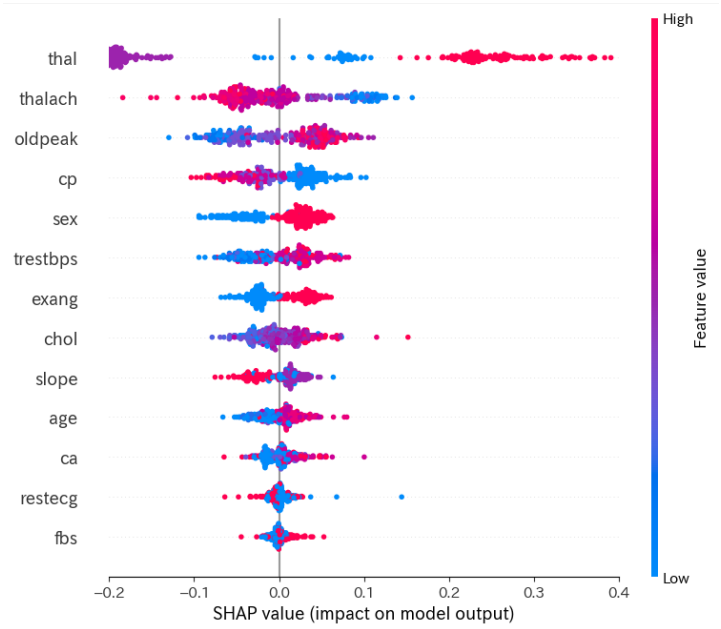

Figure 12: SHAP summary plot for KGSynX-generated synthetic data.

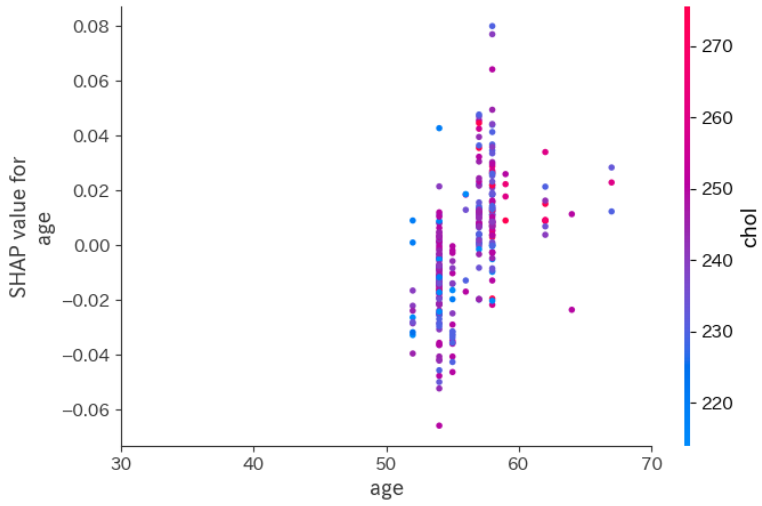

Figure 13: SHAP dependence plot for synthetic data: effect of age (colored by cholesterol).

6.1.5 統計的ギャップの概要

合成データと実データの分布の違いを定量的に評価するため、複数の整合度指標を算出した。これらのギャップ値は、個々の特徴量における偏差を示すものではなく、全属性にわたる平均値や総合的な差異を表す統計量である。これにより、平均・分散・共分散構造といった観点からデータセット全体の整合性を捉えることができ、単一の属性比較を超えた包括的な忠実度の測定を可能にする。

分析の結果、KGSynX が特徴量の順位付けや全体的な分布形状を保持する一方で、二次統計量や特徴量間の相互作用については改善の余地があることが分かった（表 8）。ここでいう特徴量の順位付けとは、グローバルな帰属階層の一貫性、すなわち実データと合成データで学習したモデルが、SHAP 値に基づき同様の特徴量重要度のランクを割り当てるかどうかを指す。これは、KGSynX が単一特徴量レベルでの周辺分布や決定ロジックを保持できていることを示唆する。

一方で、二次統計量は特徴間の分散や共分散を捉えるものであり、合成データが実データに存在する相関構造や特徴間依存をどの程度再現できるかを反映する。観測されたギャップからは、KGSynX が個々の特徴の重要性にはよく対応しているものの、属性間の結合分布や論理的制約を完全にモデル化するには、さらなる改良が必要である。

Table 8: Statistical Gaps (UCI dataset)

Metric	**Value**	**Interpretation**
Mean Gap	2.8780	Average shift in feature means
Std Deviation Gap	12.8360	Difference in spread / variability
Covariance Gap	2509.05	Difference in multivariate relationships
Spearman Rank Corr	0.7692	Rank-order consistency across features

6.2 企業向け請求書データセット

6.2.1 周辺分布と KL ダイバージェンス

分析の結果、合成データは実データの全体的な形状、特に低値領域において良好に近似していた（図 14）。しかし、有料クライアント数や使用 ID 数といった特徴量では、尾部の変動性が低下し、中央ピークが過度に平坦化する傾向が見られた。また、整合性を定量的に評価するため、実データと合成データの周辺分布間で KL ダイバージェンスを算出した（表 9）。その結果、大半の KL スコアは 0.2 未満であり、頻出値領域における良好な整合性を示していた。一方で、有料クライアント ID における 0.4072 といった一部の高いスコアは、分布の尾部における偏差を反映していると考えられる。

Table 9: KL divergence (Enterprise Invoice dataset)

Feature	**KL Divergence**
Total Invoices (6mo)	0.0519
Active Users	0.1698
Usage IDs	0.2387
Free Users	0.1698
Paid Users	-
Free Access Users	0.1665
Paid Access Users	-
Paid Clients (ID)	0.4072
Free Clients (ID)	0.0098
Issued Invoice (binary)	0.0073

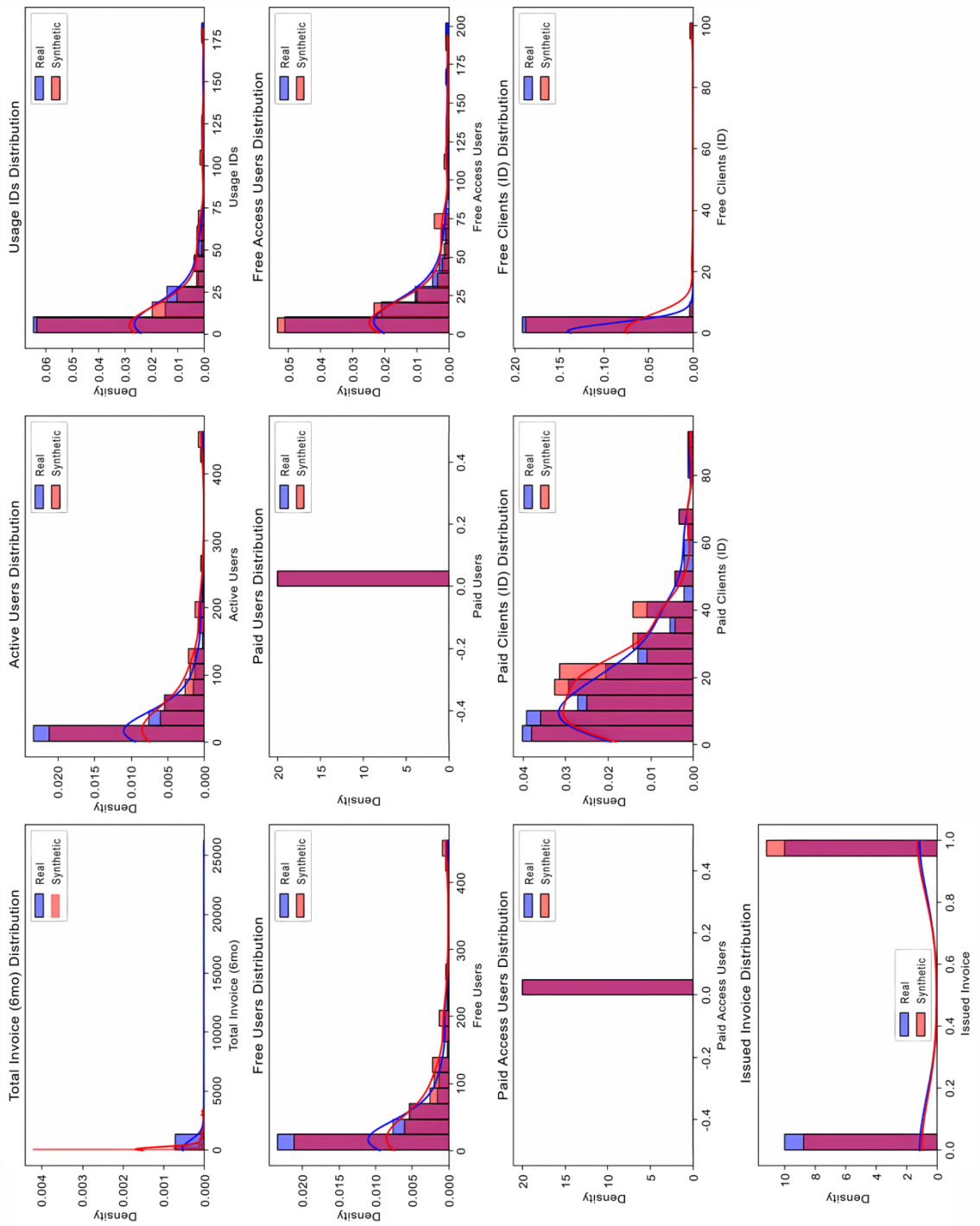

Figure 14: Density comparison (Enterprise Invoice).

6.2.2 データ構造の PCA 可視化

分析結果から、実データと合成データは中心領域で重なりを示すものの、合成サンプルはより集中する傾向が確認された（図 15）。これは稀なパターンが十

分にカバーされていないことを示唆しており、プロンプトベース生成で指摘されている既知の課題と一致する [17]。

また、実データでは第 1 主成分と第 2 主成分がそれぞれ 99.69%と 0.28%の分散を説明し、累積で 99.97%に達した（表 10）。一方で、合成データでは第 1 主成分と第 2 主成分がそれぞれ 64.84%と 13.89%であり、累積は 78.73%にとどまった。この違いは、実データにおける分散が複数の方向に分散している一方で、KGSynX による合成データの変動性が主に第 1 成分に集約されていることを示している。すなわち、合成データはより緊密にクラスタリングされる傾向にあり、今後の生成プロセスでは多方向の分散を適切に保持する仕組みが求められる。

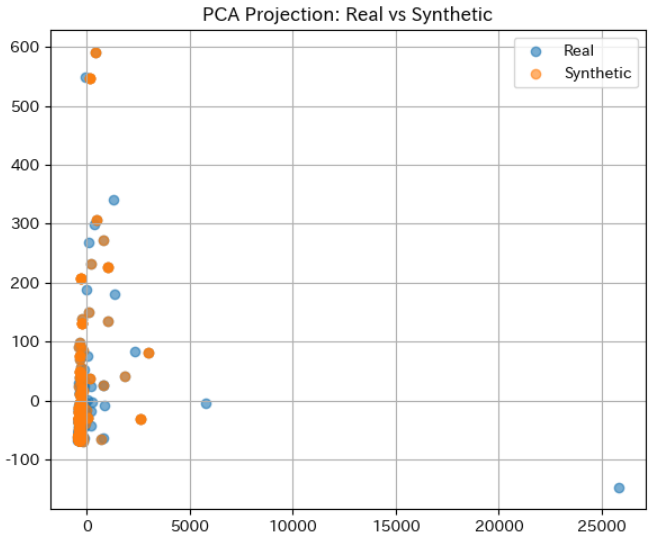

Figure 15: PCA projection (Enterprise Invoice).

Table 10: Explained variance ratio of the first two principal components (Enterprise Invoice)

Dataset	PC1 (%)	PC2 (%)	PC1+PC2 (%)
Real data	99.69	0.28	99.97
Synthetic data	64.84	13.89	78.73

6.2.3 SHAP に基づく特徴量帰属の一貫性分析

分析の結果、クライアント ID や使用回数といった最も重要な特徴量は、いずれのモデルにおいても上位となった。一方で、「無料アクセスユーザー」や二値変数「請求書発行」といった中程度の重要度をもつ特徴量については、両モデル

間で若干の順位のずれが見られた。これはサンプリングバイアスやプロンプト設計の簡略化に起因する可能性がある。

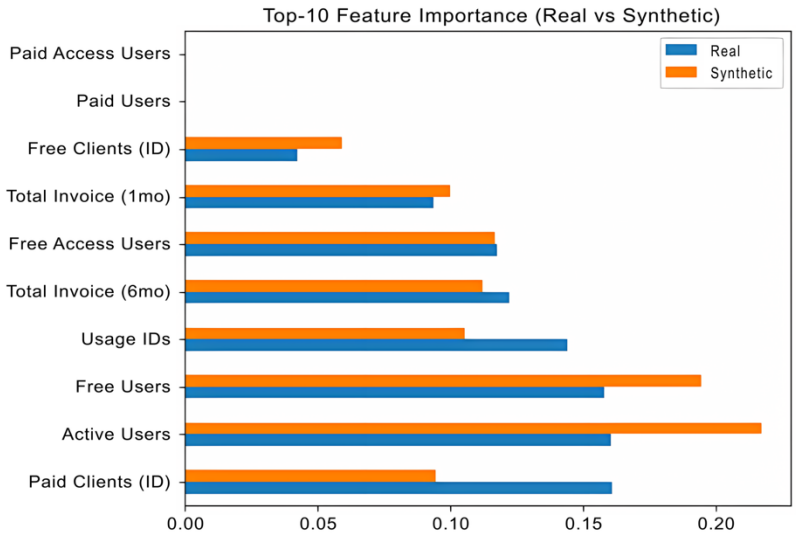

Figure 16: SHAP feature attribution comparison (Enterprise Invoice).

6.2.4 SHAP に基づく特徴量帰属の一貫性分析

実データに基づく SHAP サマリープロットの分析からは、NumUsers、NumIDs、NumFreeUsers といったアクティブな利用指標が予測モデルにおける主要な寄与因子であることが確認された（図 17）。これらの特徴量は一貫した方向性をもつ影響を示しており、利用量が増えるほど請求書発行の可能性が高まるという明確な相関関係が認められる。

さらに SHAP 依存性プロットの分析では、InvoiceCount_1Month と InvoiceCount_6MonthTotal の間に非線形ではあるが解釈可能な相互作用が存在することが分かった（図 18）。具体的には、SHAP 値は一定の閾値を超えると急激に上昇し、その後は飽和状態に達する傾向が観察された。

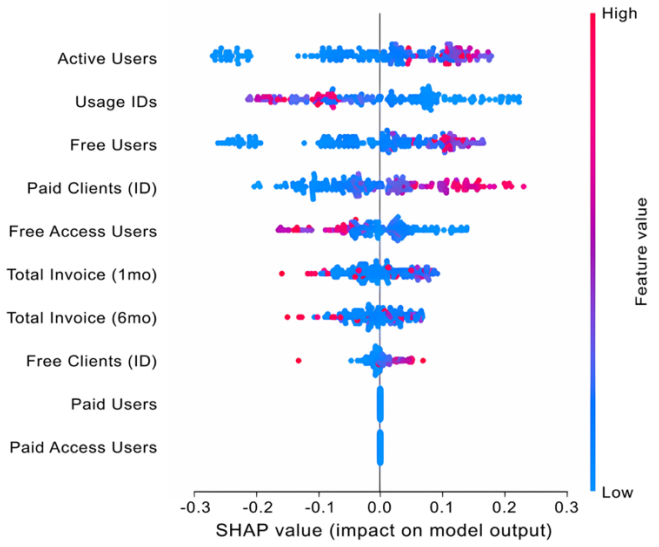

Figure 17: SHAP summary plot for real Infomart data.

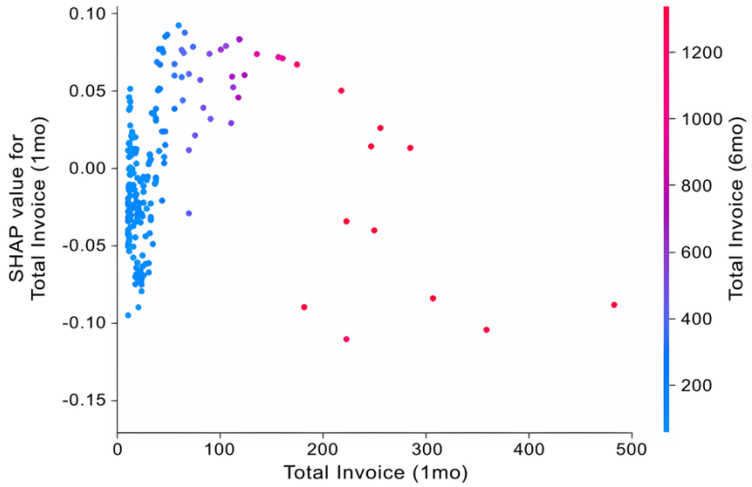

Figure 18: SHAP dependence plot for real data: effect of recent monthly invoices (colored by 6-month cumulative count).

合成データに注目すると、KGSynX は特徴量帰属の高次構造を概ね保持していることが分かる（図 19）。合成モデルにおける上位特徴量は実データモデルとほぼ一致しており、その SHAP 値も同様である。例えば、NumUsers や NumFreeUsers は依然として請求書発行行動の強力な予測因子として残っている。依存性プロットの結果も、請求書活動と SHAP の影響力の間に単調なパターンを示しており、学習された相互作用ロジックが保持されていることを示唆している（図 20）。

一方で、いくつかの重要な相違点も観察された。第一に、特徴量帰属の階層が合成データでは平坦化しており、主要特徴量と副次的特徴量の差異が弱まっている。第二に、SHAP 値の分布幅が広がり、インスタンス間で帰属強度のばらつきが大きくなっている。これはプロンプトベースの生成過程で導入される確率的変動の影響と考えられる。最後に、合成データの依存性プロットでは、相互作用パターンがより滑らかで対称的に現れ、実データで見られる飽和効果が確認されなかった。これらの観察結果は、極端な値のモデリングや特徴間の依存関係について、さらなる改善の余地があることを示している。

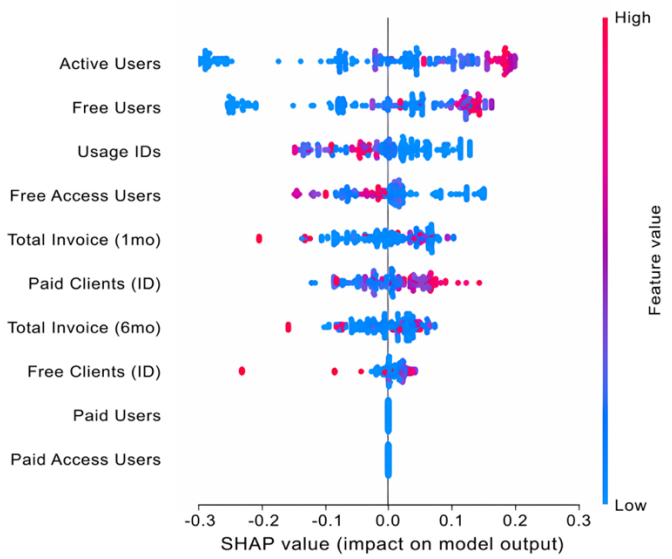

Figure 19: SHAP summary plot for KGSynX-generated synthetic data (Infomart).

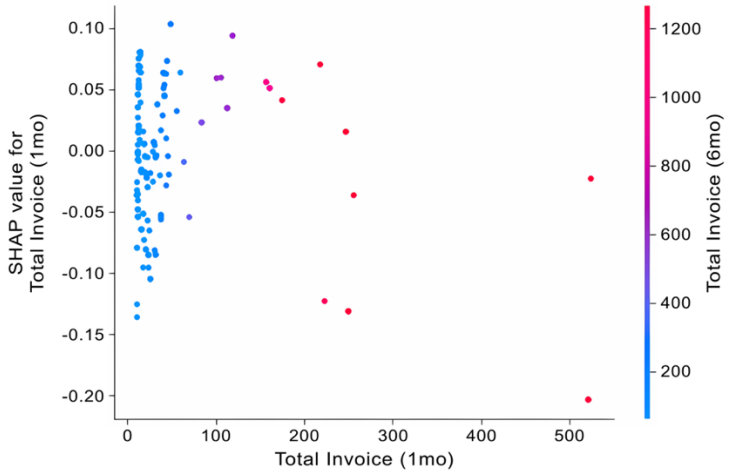

Figure 20: SHAP dependence plot for synthetic data: effect of recent monthly invoices (colored by 6-month cumulative count).

6.2.5 統計的ギャップの概要

全ての分布差異指標を集約した結果、比較的大きな平均値ギャップと共分散ギャップが確認された（表 11）。これは、一変量レベルでは良好な整合性が得られているにもかかわらず、高次の相互作用を十分に捉えきれていない課題を浮き彫りにしている。特に、本データにおける共分散ギャップは、他の 2 つのデータセットに比べて非常に大きい。この現象には主に 2 つの要因が考えられる。第一に、請求書データセットの特徴量値は数桁にわたる大きなオーダーを持つ一方で、Heart Disease データの生理学的測定値は比較的狭い範囲に収まり、より均一なスケールをもつ。事前の正規化やスケーリングを行わずに算出された共分散行列は大きな特徴量に強く影響を受けるため、請求書データの場合にはフロベニウスノルムにおける不一致が大きくなる。第二に、請求書データの本来の共分散構造は本質的に密であり、生成モデルはこうした複雑な特徴量相互作用を再現しようとする際に共分散情報を失いやすい傾向があると考えられる。

Table 11: Statistical Gaps (Enterprise Invoice Dataset)

Metric	**Value**
Mean Gap	19.6455
Standard Deviation Gap	143.5829
Covariance Gap	3,466,926.39
Spearman Rank Corr.	0.6951

6.3 通信事業者の顧客離反データ

6.3.1 周辺分布と KL ダイバージェンス

4 つの主要な連続変数（SeniorCitizen、tenure、MonthlyCharges、TotalCharges）について、実データと合成データの周辺分布を比較した。二値変数である SeniorCitizen は、実データ・合成データの双方で明確な二分化を示し、合成分布は 2 つのピークをほぼ再現していた（表 12）。これはカテゴリ分布が正確に保持されていることを示している。顧客の契約月数を表す tenure は、実データでは複雑な多峰性をもち、契約ライフサイクルのパターンを反映していた。合成データは主要なモードを捉えているが、中間的な頻度スパイクの一部は平滑化されていた。具体的には、典型的な顧客在籍期間（1 か月目、12 か月目、24 か月目、60 か月目など）は再現されているものの、中頻度の顧客セグメントは十分に表現されていないことが示唆される。

MonthlyCharges と TotalCharges については、分布の大部分で良好な近似が見られたが、上位範囲では密度がやや低下していた（図 21）。特に高額支出の長期顧客に相当する極端値は過小評価される傾向にあり、これはプロンプトベース生成でしばしば指摘される問題と一致している。また、KL ダイバージェンスの結果もこの傾向を裏付けている。全てのスコアは 0.3 未満であり、周辺統計量レベルでは非常に強い一致が確認された。中でも SeniorCitizen の極めて低い KL 値（0.0032）は、モデルが二値構造を正しく捉えていることを示す。一方で、MonthlyCharges のやや高い KL 値は、そのロングテール特性が合成データで十分に再現されていないことを反映している。

Table 12: KL divergence (Telco Churn dataset)

Feature	**KL Divergence**
SeniorCitizen	0.0032
Tenure	0.1616
MonthlyCharges	0.2991
TotalCharges	0.1227

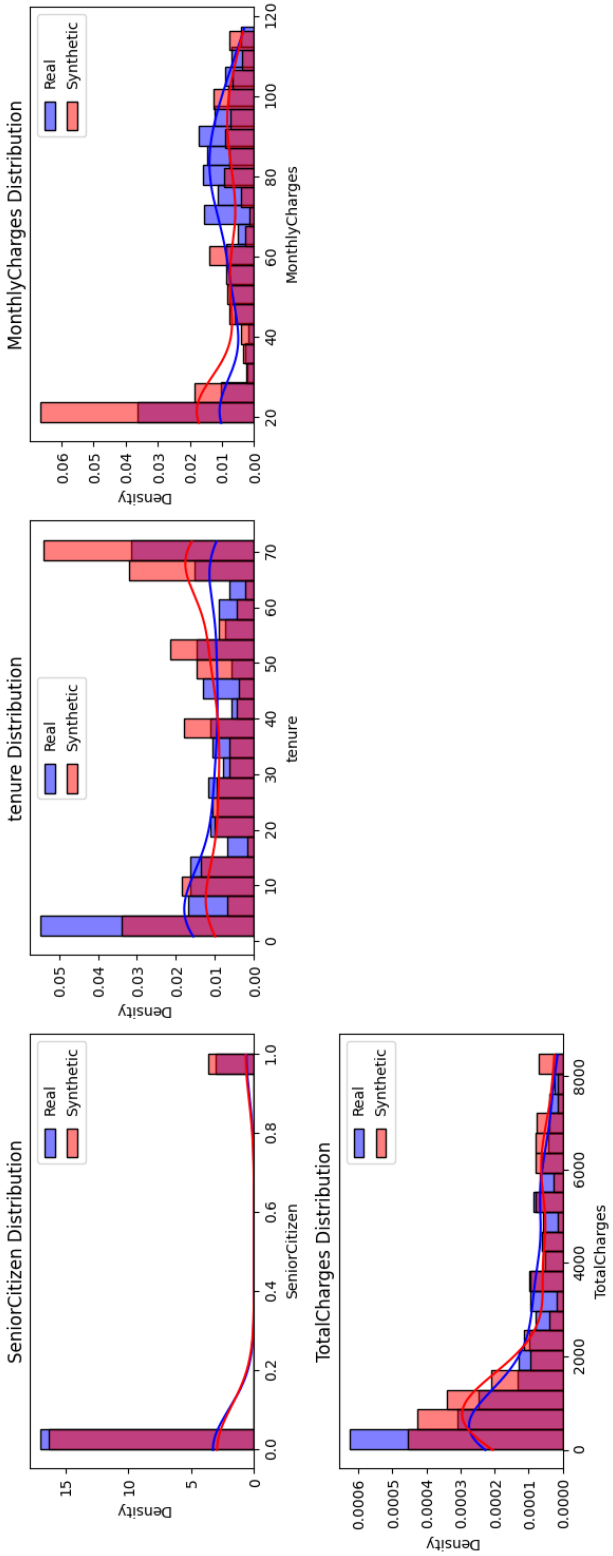

Figure 21: Density comparison (Telco).

6.3.2 データ構造の PCA 可視化

PCA 可視化の結果、実データと合成データは顕著な重なりを示し、両者のクラスターは類似した形状と曲率を持つことが確認された（図 22）。請求書利用状況データでは合成点が中心に密集する傾向が見られたのに対し、ここでは合成データがより広い範囲に分布し、中心への偏りが少ない。これは、KGSynX が本データセットにおいて全体的な分布トポロジーをより効果的に保持していることを示しており、その背景にはより明確なスキーマ構造やバランスの取れた特徴空間の存在があると考えられる。

中心領域の密度が高いことは、モデルが典型的な顧客行動パターンを適切に学習していることを示唆している。しかし一方で、図の端部、すなわち稀なサービスの組み合わせや極端に高額／低額の請求に対応する領域では、合成データにおいても疎な分布となっており、尾部のモデリングに関する課題が依然として残っていることが分かる。

この観察を定量的に検証するため、最初の 2 つの主成分が説明する分散比率を比較した（表 13）。実データでは PC1 が 99.99%を占め、PC2 はわずか 0.01%であり、ほぼ一次元的な支配パターンを反映していた。これに対して、合成データでは PC1 が 64.84%、PC2 が 13.89%（合計 78.73%）を説明しており、主要な分散方向は再現されているものの、副次的成分に沿った変動性は過小評価されていた。これは、PCA で見られた尾部カバレッジの不足とも一致している。

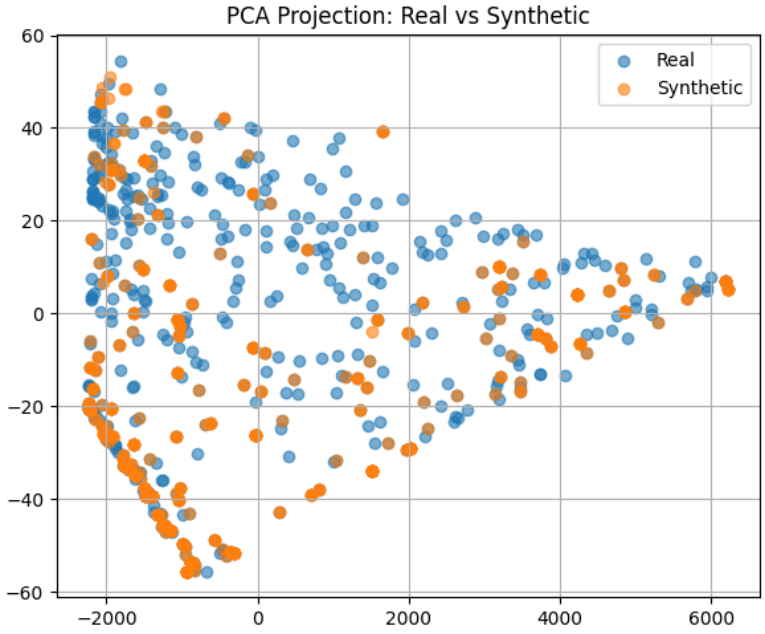

Figure 22: PCA projection (Telco).

Table 13: Explained variance ratio of the first two principal components (Telco)

Dataset	PC1 (%)	PC2 (%)	PC1+PC2 (%)
Real data	99.99	0.01	100.00
Synthetic data	64.84	13.89	78.73

6.3.3 SHAP アトリビューションの整合性

実データと合成データでそれぞれ訓練した分類器に対し、SHAP 値で上位 10 の特徴量を比較した結果、契約、契約期間、総料金といった主要な予測変数は、両モデルにおいて一貫して支配的な特徴量として特定された（図 23）。興味深い点として、合成データで訓練したモデルは Contract を特に重視しており、その寄与度は約 40%に達していた。これは実データモデルで見られる特徴重要度の分散とは対照的である。この偏りは、プロンプト設計において契約タイプがテンプレートに明示的に含まれていることに起因すると考えられる。その結果、モデルは契約タイプの論理を過度に学習し、月間請求額や支払い方法などの課金関連特徴量の相対的な重要度が低下していた。このような帰属の偏りは分類精度には影響しないものの、解釈可能性や推論の柔軟性を損なう可能性がある。一方で、TotalCharges、Partner、OnlineSecurity などの特徴量における整合性は、顧客属性や人口統計的要因の関係性が適切に保持されていることを示している。

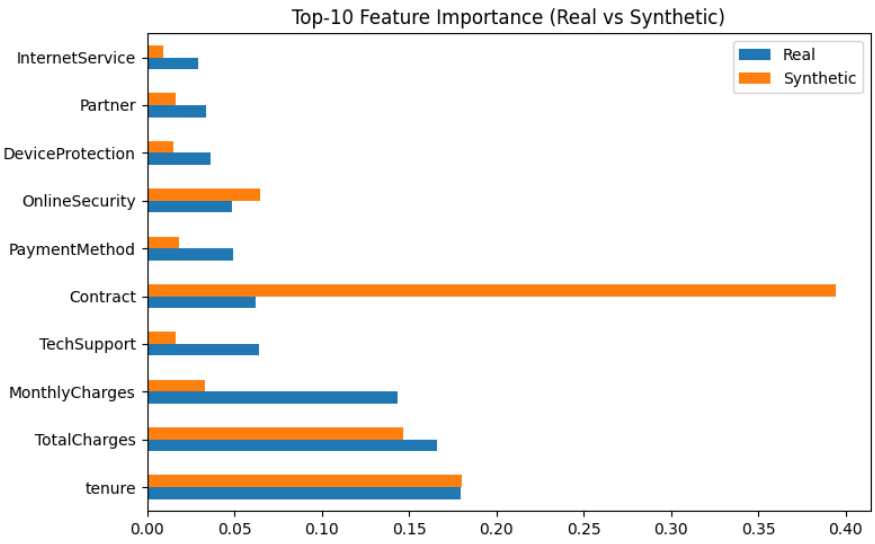

Figure 23: SHAP feature attribution comparison (Telco).

6.3.4 SHAP に基づく特徴帰属の一貫性分析

実データに基づく SHAP のサマリープロットと依存性プロットを分析した結果、tenure、MonthlyCharges、TotalCharges がモデル挙動を強く支配していることが明らかとなった（図 24）。これは既に知られている顧客解約のダイナミクスと整合している。一般的に、tenure と TotalCharges の値が高いほど解約確率は低下する傾向にある。一方、MonthlyCharges についてはより複雑な相互作用が見られ、単純な線形関係では説明できない振る舞いを示していた。

依存性プロットの結果からは、MonthlyCharges や tenure といった二次的な変数を条件にした場合に、顧客の契約期間や累積課金額がモデルの出力にどのように影響するかが明確に示されている（図 25）。これにより、モデルが解約リスクを予測する際に、単一の特徴量だけでなく、複数の属性の組み合わせを踏まえた意思決定を行っていることが確認できる。

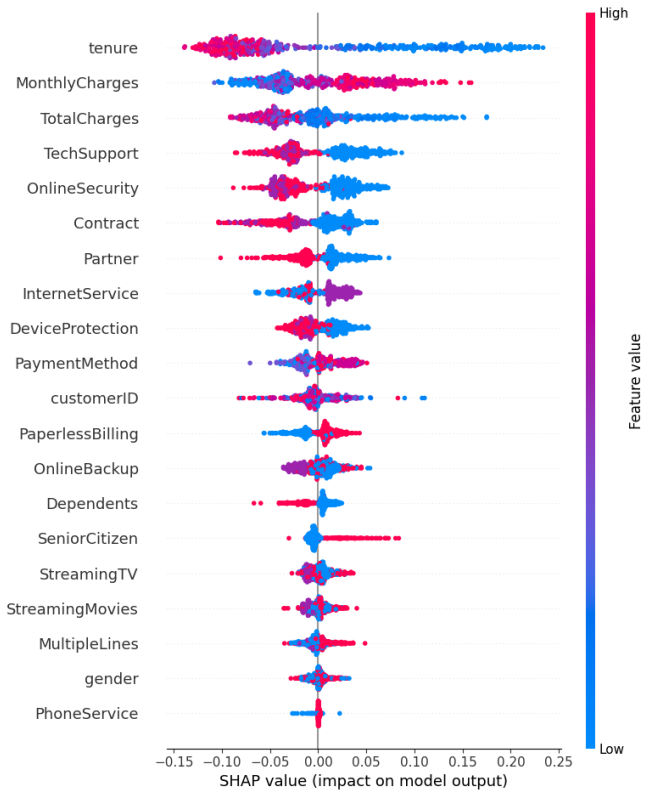

Figure 24: SHAP summary plot for real Telco Customer Churn data.

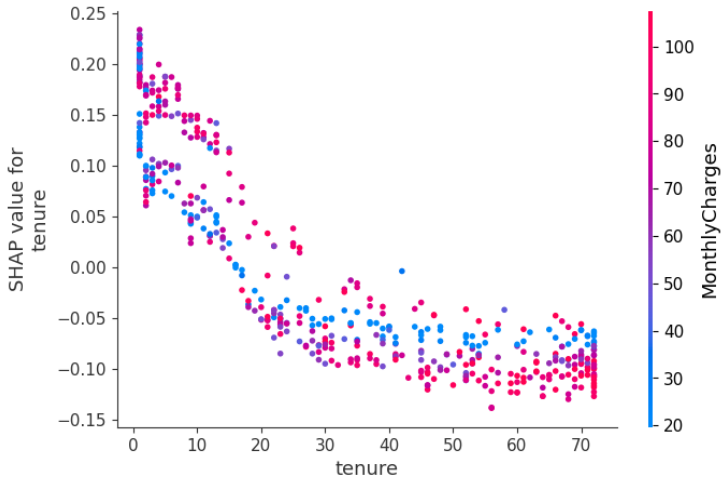

Figure 25: SHAP dependence plot for real data: effect of TotalCharges (colored by tenure).

合成データに基づく SHAP 分析の結果、上位の特徴量は概ね一貫して維持されており、依存性プロットも実データと類似した傾向を示していた（図 26）。特に注目すべき点は、tenure と解約リスクの関係が保持されており、契約初期段階に見られる強い非線形性も再現されていたことである（図 27）。

一方で、いくつかの差異も観察された。まず、合成データの SHAP サマリープロットでは特徴重要度の分布がやや平坦化しており、契約期間やオンラインセキュリティといった上位特徴量間での重要度の差が明確に現れにくくなっていた（図 28）。さらに、SHAP 値のばらつきが減少しており、生成過程における過剰な正則化の影響が示唆される。最後に、依存性プロットにおける相互作用効果は、特に分布の尾部で滑らかに表現される傾向があり、KGSynX が外れ値パターンを過小評価する可能性があることが示された。

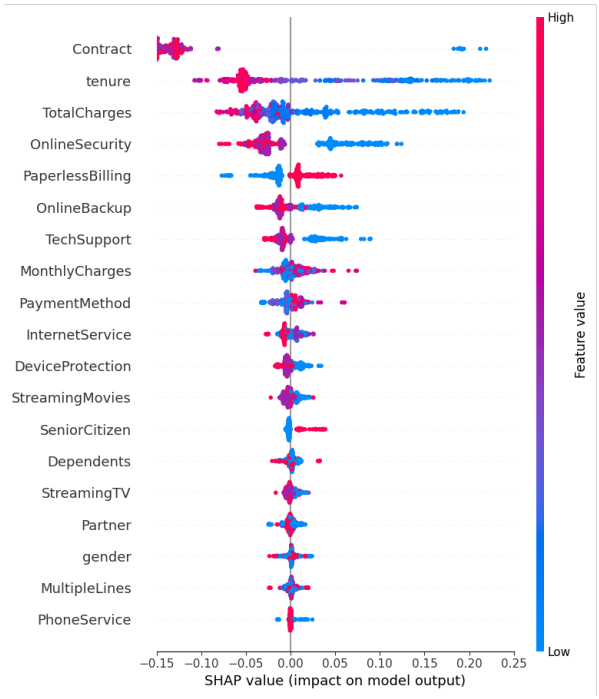

Figure 26: SHAP summary plot for KGSynX-generated synthetic Telco data.

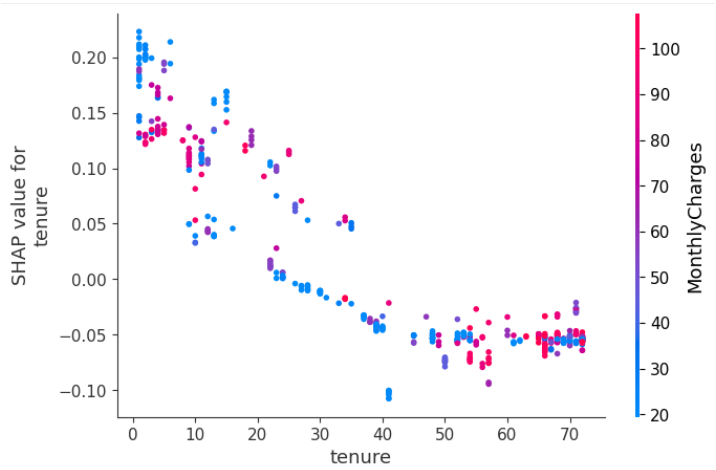

Figure 27: SHAP dependence plot for synthetic data: effect of tenure (colored by MonthlyCharges).

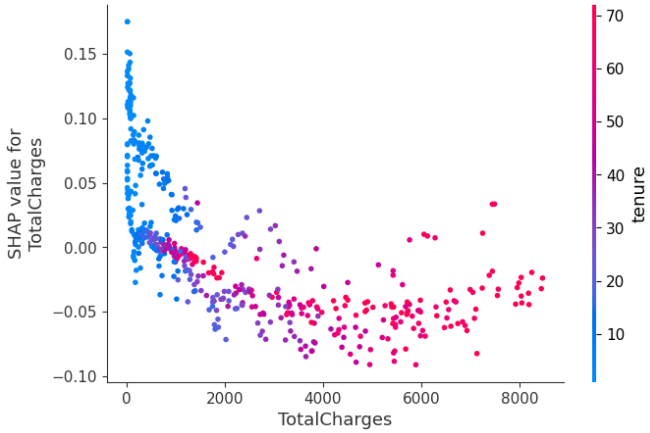

Figure 28: SHAP dependence plot for synthetic data: effect of TotalCharges (colored by tenure).

6.3.5 統計的ギャップの概要

分析の結果、3 つのデータセットの中で最も高い意味的整合性が達成されていることが確認された（表 14）。これは、最も低い共分散ギャップと最高のスピアマン順位相関係数（0.8702）を示しており、特徴間の論理構造や特徴の順位付けが良く保持されていることを意味している。一方で、平均ギャップは比較的高い値を示しており、これは TotalCharges のような連続変数の影響を受けた結果であると考えられる。しかし、全体的な分布構造は依然として良好に維持されている。

UCI データセットや Enterprise データセットと比較すると、この結果は、KGSynX が特徴の粒度がバランスされ、属性の意味論が比較的単純でありながら、予測タスク（例えば顧客解約の分類）に強く関連する領域で、最も優れた性能を発揮することを示唆している。

Table 14: Statistical Gaps (Telco)

Metric	**Value**
Mean Gap	24.9491
Standard Deviation Gap	30.7742
Covariance Gap	558,392.03
Spearman Rank Corr.	0.8702

7. 結論

本論文では、知識グラフによる構造モデリング、プロンプト誘導型 LLM 生成、SHAP に基づくフィードバック改良を統合し、統計的整合性、構造的意味、予測に対する解釈可能性を同時に保持する合成表形式データ生成手法 KGSynX を提案した。

本研究の貢献は以下に集約される。第 1 に、表形式データをインスタンスレベルとスキーマレベルの関係性を保持し、知識グラフへと変換する手法を提案した点である。これにより、従来の平坦な特徴量ベースの生成では保持が難しいドメイン論理を明示的に扱えるようになった。第 2 に、グラフ埋め込みを用いて意味的情報量の多いプロンプトを構築した点である。これにより、タスク特化の再学習を行わずに構造的知識を反映した生成を可能にした。第 3 に、SHAP を導入し、実データと合成データで学習したモデルの間に生じる意味的ギャップを定量化・修正できる仕組みを実現した点である。

実験では、UCI Heart Disease、Enterprise Invoice Usage、Telco Customer Churn の 3 つの実データセットを用い、KGSynX の有効性を検証した。その結果、分類性能と SHAP ベースの意味的整合性の両面において、既存手法である CTGAN や単純な LLM プロンプティングなどを大きく上回った。さらに、KL ダイバージェンスや順位相関などの分布指標でも有用な結果を示し、下流タスクにおける実用性が認められる結果を得た。

もちろん、KGSynX にはいくつかの限界が存在する。第 1 に、厳格な論理制約を直接課さず、ソフトな知識表現と反復的修正に依存しているため、安全性が極めて重要な分野では適用が制限される可能性がある。第 2 に、プロンプト生成を部分的に手作業で行う必要があり、あらゆる分野に汎用的に適用できる設計になっていない。第 3 に、SHAP 計算への依存が計算コストを増大させ得る点である。これらは実用化における重要な課題となる可能性が高い。

今後の展望としては、知識グラフによるソフトな制約を補うための形式論理エンジンや制約ソルバーの統合、強化学習を用いた自動かつ文脈に応じたプロンプト最適化の導入が考えられる。また、LIME や統合勾配法、反実仮想といった他の解釈手法をフィードバックループに組み込むことで、より多様なモデルや状況に適応できるかもしれない。さらに、実環境での展開や専門家を交えた評価を行うことで、生成データの信頼性と現場での受容性を高める方法も考えられる。

参考文献

[1] Prabal Banerjee and Sushmita Ruj. Blockchain enabled data marketplace–design and challenges. arXiv preprint arXiv:1811.11462, 2018.

[2] Mohammad Rasouli and Michael I Jordan. Data sharing markets. arXiv preprint arXiv:2107.08630, 2021.

[3] Trivellore E Raghunathan. Synthetic data. Annual review of statistics and its application, 8:129–140, 2021.

[4] Emiliano De Cristofaro. Synthetic data: Methods, use cases, and risks. arXiv preprint arXiv:2303.01230, 2024.

[5] Tshilidzi Marwala, Eleonore Fournier-Tombs, and Serge Stinckwich. The use of synthetic data to train ai models: Opportunities and risks for sustainable development, arXiv preprint arXiv:2309.00652, 2023.

[6] Joao Fonseca and Fernando Bacao. Tabular and latent space synthetic data generation: a literature review. Journal of Big Data, 10:115, 2023.

[7] Paul Voigt and Axel von dem Bussche. The EU General Data Protection Regulation (GDPR). Springer, 2017.

[8] Cynthia Dwork. Differential privacy: A survey of results. Theory and Applications of Models of Computation, 2008.

[9] Andre Goncalves, Priyadip Ray, Braden Soper, Jennifer Stevens, Linda Coyle, and Ana Paula Sales. Generation and evaluation of synthetic patient data. BMC Medical Research Methodology, 20:108, 2020.

[10] Mauro Giuffrè and Dennis L. Shung. Harnessing the power of synthetic data in healthcare: innovation, application, and privacy. NPJ digital medicine, 6:186, 2023.

[11] Neha Patki, Roy Wedge, and Kalyan Veeramachaneni. The synthetic data vault. International Conference on Data Science and Advanced Analytics, 2016.

[12] Snowflake: Using synthetic data in snowflake. https://www.snowflake.com/blog/, 2025.

[13] Ramiro Camino, Christian Hammerschmidt, Radu State. Generating multi-categorical samples with generative adversarial networks. arXiv preprint arXiv:1807.01202, 2018.

[14] Lei Xu, Maria Skoularidou, Alfredo Cuesta-Infante, and Kalyan Veeramachaneni. Modeling tabular data using conditional GAN. Advances in Neural Information Processing Systems, 2019.

[15] Scott M Lundberg and Su-In Lee. A unified approach to interpreting model predictions. Advances in Neural Information Processing Systems, 2017.

[16] Aidan Hogan, Eva Blomqvist, Michael Cochez, Claudia D'amato, Gerard De Melo, Claudio Gutierrez, Sabrina Kirrane, José Emilio Labra Gayo, Roberto Navigli, Sebastian Neumaier, Axel-Cyrille Ngonga Ngomo, Axel Polleres, Sabbir M. Rashid, Anisa Rula, Lukas Schmelzeisen, Juan Sequeda, Steffen Staab, and Antoine Zimmermann. Knowledge graphs. ACM Computing Surveys, 54(4):1–37, 2021.

[17] Jinhao Jiang, Kun Zhou, Zican Dong, Keming Ye, Wayne Xin Zhao, and Ji-Rong Wen. StructGPT: A general framework for large language model to reason over structured data. arXiv preprint arXiv:2305.09645, 2023.

[18] Richard J. Chen, Ming Y. Lu, Tiffany Y. Chen, Drew F. K. Williamson, and Faisal Mahmood. Synthetic data in machine learning for medicine and healthcare. Nature Biomedical Engineering, 5:493–497, 2021.

[19] Vadim Borisov, Tobias Leemann, Kathrin Seßler, Johannes Haug, Martin Pawelczyk, Gjergji Kasneci. Deep neural networks and tabular data: A survey. IEEE Transactions on Neural Networks and Learning Systems, 35(6):7499-7519, 2024.

[20] Aristidis K. Nikoloulopoulos and Dimitris Karlis. Modeling multivariate count data using copulas. Communications in Statistics-Simulation and Computation, 39(1):172–187, 2009.

[21] Regis Houssou, Mihai-Cezar Augustin, Efstratios Rappos, Vivien Bonvin, Stephan Robert-Nicoud. Generation and simulation of synthetic datasets with copulas. arXiv preprint arXiv:2203.17250, 2022.

[22] Marco Ramoni and Paola Sebastiani. Learning Bayesian networks from incomplete databases. 13th conference on Uncertainty in artificial intelligence, 1997.

[23] Larissa N. A. Martins, Flávio B. Gonçalves, and Thais P. Galletti. Generation and analysis of synthetic data via Bayesian networks: a robust approach for uncertainty quantification via Bayesian paradigm, arXiv preprint arXiv:2402.17915, 2024.

[24] Edward Choi, Siddharth Biswal, Bradley Malin, Jon Duke, Walter F. Stewart, Jimeng Sun. Generating multi-label discrete patient records using generative adversarial networks. 2nd Machine Learning for Healthcare Conference, 2017.

[25] Akim Kotelnikov, Dmitry Baranchuk, Ivan Rubachev, Artem Babenko. TabDDPM: Modelling Tabular Data with Diffusion Models. arXiv preprint arXiv:2209.15421, 2023.

[26] Sierra Wyllie, Ilia Shumailov, and Nicolas Papernot. Fairness feedback loops: training on synthetic data amplifies bias. ACM Conference on Fairness, Accountability, and Transparency, 2024.

[27] OpenAI, Josh Achiam, Steven Adler, Sandhini Agarwal et al. GPT-4 technical report, 2024.

[28] Tom B. Brown, Benjamin Mann, Nick Ryder, et al. Language models are few-shot learners. NeurIPS, 2020.

[29] Joy Mahapatra and Utpal Garain. An extensive evaluation of factual consistency in large language models for data-to-text generation. arXiv preprint arXiv:2411.19203, 2024.

[30] Allison Koenecke, Andrew Nam, Emily Lake, Joe Nudell, Minnie Quartey, Zion Mengesha, Connor Toups, John R. Rickford, Dan Jurafsky, and Sharad Goel. Racial disparities in automated speech recognition. Proceedings of the National Academy of Sciences, 117(14):7684–7689, 2020.

[31] Yingzhou Lu, Lulu Chen, Yuanyuan Zhang, Minjie Shen, Huazheng Wang, Xiao Wang, Capucine van Rechem, Tianfan Fu, Wenqi Wei. Machine learning for synthetic data generation: a review. arXiv preprint arXiv:2302.04062, 2023.

[32] Shaoxiong Ji, Shirui Pan, Erik Cambria, Pekka Marttinen, and Philip S. Yu. A survey on knowledge graphs: Representation, acquisition, and applications. IEEE Transactions on Neural Networks and Learning Systems, 33(2):494–514, 2022.

[33] Zongbao Yang, Yuchen Lin, Yinxin Xu, Jinlong Hu, and Shoubin Dong. Interpretable disease prediction via path reasoning over medical knowledge graphs and admission history. Knowledge-Based Systems, 281:111082, 2023.

[34] Heiko Paulheim. Knowledge graph refinement: A survey of approaches and evaluation methods. Semantic Web, 8(3):489–508, 2016.

[35] Xiaozhi Wang, Tianyu Gao, Zhaocheng Zhu, Zhengyan Zhang, Zhiyuan Liu, Juanzi Li, Jian Tang. Kepler: A unified model for knowledge embedding and pre-trained language representation. arXiv preprint arXiv:1911.06136, 2019.

[36] Zhengyan Zhang, Xu Han, Zhiyuan Liu, Xin Jiang, Maosong Sun, and Qun Liu. Ernie: Enhanced language representation with informative entities. 57th Annual Meeting of the Association for Computational Linguistics, 2019.

[37] Maxat Kulmanov, Fatima Zohra Smaili, Xin Gao, Robert Hoehndorf. Machine learning with biomedical ontologies. bioRxiv 2020.05.07.082164, 2020.

[38] Antoine Bordes, Nicolas Usunier, Alberto Garcia-Duran, Jason Weston, and Oksana Yakhnenko. Translating embeddings for modeling multi-relational data. NeurIPS, 2013.

[39] Bishan Yang, Wen-tau Yih, Xiaodong He, Jianfeng Gao, and Li Deng. Embedding entities and relations for learning and inference in knowledge bases. arXiv preprint arXiv:1412.6575, 2014.

[40] Aditya Grover and Jure Leskovec. node2vec: Scalable feature learning for networks. ACM SIGKDD, 2016.

[41] Quan Wang, Zhendong Mao, Bin Wang, and Li Guo. Knowledge graph embedding: A survey of approaches and applications. IEEE Transactions on Knowledge and Data Engineering, 29(12):2724–2743, 2017.

[42] Xi Fang, Weijie Xu, Fiona Anting Tan, Jiani Zhang, Ziqing Hu, Yanjun Qi, Scott Nickleach, Diego Socolinsky, Srinivasan Sengamedu, Christos Faloutsos. Large language models (LLMs) on tabular data: Prediction, generation, and understanding–a survey. arXiv preprint arXiv:2402.17944, 2024.

[43] Pengcheng Jiang, Lang Cao, Ruike Zhu, Minhao Jiang, Yunyi Zhang, Jimeng Sun, Jiawei Han. RAS: Retrieval-And-Structuring for Knowledge-Intensive LLM Generation. arXiv preprint arXiv:2502.10996, 2025.

[44] Yujia Qin, Shihao Liang, Yining Ye, Kunlun Zhu, Lan Yan, Yaxi Lu, Yankai Lin, Xin Cong, Xiangru Tang, Bill Qian, Sihan Zhao, Lauren Hong, Runchu Tian, Ruobing Xie, Jie Zhou, Mark Gerstein, Dahai Li, Zhiyuan Liu, Maosong Sun. ToolLLM: Facilitating Large Language Models to Master 16000+ Real-world APIs. arXiv preprint arXiv:2307.16789, 2023.

[45] Baolin Peng, Michel Galley, Pengcheng He, Hao Cheng, Yujia Xie, Yu Hu, Qiuyuan Huang, Lars Liden, Zhou Yu, Weizhu Chen, Jianfeng Gao. Check Your Facts and Try Again: Improving Large Language Models with External Knowledge and Automated Feedback. arXiv preprint arXiv:2302.12813, 2023.

[46] Finale Doshi-Velez and Been Kim. Towards a rigorous science of interpretable machine learning. arXiv preprint arXiv:1702.08608, 2017.

[47] Riccardo Guidotti, Anna Monreale, Salvatore Ruggieri, et al. A survey of methods for explaining black box models. ACM Computing Surveys, 51(5):1–42, 2018.

[48] Rudresh Dwivedi, Devam Dave, Het Naik, Smiti Singhal, Rana Omer, Pankesh Patel, Bin Qian, Zhenyu Wen, Tejal Shah, Graham Morgan, Rajiv Ranjan. Explainable AI (XAI): Core ideas, techniques, and solutions. ACM computing surveys, 55(9):1–33, 2023.

[49] Scott M. Lundberg, Bala Nair, Monica S. Vavilala, Mayumi Horibe, Michael J. Eisses, Trevor Adams, David E. Liston, Daniel King-Wai Low, Shu-Fang Newman, Jerry Kim, Su-In Lee. Explainable machine-learning predictions for the prevention of hypoxaemia during surgery. Nature Biomedical Engineering, 2:749–760, 2018.

[50] Lara Marie Demajo, Vince Vella, and Alexiei Dingli. Explainable AI for interpretable credit scoring. arXiv preprint arXiv:2012.03749, 2020.

[51] Jitendra Maan and Harsh Maan. Customer churn prediction model using explainable machine learning. arXiv preprint arXiv:2303.00960, 2023.

[52] Shibani Santurkar, Ludwig Schmidt, and Aleksander Madry. A Classification-Based Study of Covariate Shift in GAN Distributions. International Conference on Machine Learning, 2018.

[53] Chance N. DeSmet and Diane J. Cook. Hydragan: A multi-head, multi-objective approach to synthetic data generation. arXiv preprint arXiv:2111.07015, 2021.

[54] Jinsung Yoon, Logan Drumright, and Mihaela Van Der Schaar. Anonymization through data synthesis using generative adversarial networks (ADS-GAN). IEEE Journal of Biomedical and Health Informatics, 24(8):2378–2388, 2020.

[55] Alexis Fox, Samarth Swarup, and Abhijin Adiga. A unifying information- theoretic perspective on evaluating generative models. AAAI Conference on Artificial Intelligence, 2025.

[56] Dayananda Herurkar, Ahmad Ali, and Andreas Dengel. Evaluating generative models for tabular data: Novel metrics and benchmarking. arXiv preprint arXiv:2504.20900, 2025.

[57] Pengfei Liu, Weizhe Yuan, Jinlan Fu, Zhengbao Jiang, Hiroaki Hayashi, and Graham Neubig. Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing. ACM computing surveys, 55(9):1–35, 2023.

[58] Wenhu Chen. Large language models are few (1)-shot table reasoners. EACL, 2023.

[59] Stefan Hegselmann, Alejandro Buendia, Hunter Lang, Monica Agrawal, Xiaoyi Jiang, and David Sontag. TabLLM: Few-shot classification of tabular data with large language models. International Conference on Artificial Intelligence and Statistics, 2023.

[60] Lauren Arthur, Jason Costello, Jonathan Hardy, Will O'Brien, James Rea, Gareth Rees, and Georgi Ganev. On the challenges of deploying privacy- preserving synthetic data in the enterprise. arXiv preprint arXiv:2307.04208, 2023.

[61] András Janosi, William Steinbrunn, Matthias Pfisterer, Robert Detrano. Heart disease dataset. 1989. UCI Machine Learning Repository, https://archive.ics.uci.edu/ml/datasets/Heart+Disease.

[62] Karim Armanious, Chenming Jiang, Marc Fischer, Thomas Küstner, Tobias Hepp, Konstantin Nikolaou, Sergios Gatidis, and Bin Yang. MedGAN: Medical image translation using GANs. Computerized medical imaging and graphics, 79:101684, 2020.

[63] James M. Joyce. Kullback-Leibler Divergence. International encyclopedia of statistical science, pages 720–722. Springer, 2011.

[64] Ilya Tolstikhin, Olivier Bousquet, Sylvain Gelly, and Bernhard Schölkopf. Wasserstein auto-encoders. arXiv preprint arXiv:1711.01558, 2017.

[65] Leo Breiman. Random forests. Machine learning, 45:5–32, 2001.
[66] Tianqi Chen, Carlos Guestrin. Xgboost: A scalable tree boosting system. ACM SIGKDD, 2016.
[67] Guolin Ke, Qi Meng, Thomas Finley, Taifeng Wang, Wei Chen, Weidong Ma, Qiwei Ye, Tie-Yan Liu. LightGBM: a highly efficient gradient boosting decision tree. NeurIPS, 2017.
[68] Ian T. Jolliffe and Jorge Cadima. Principal Component Analysis: a review and recent developments. Philosophical Transactions of the Royal Society A, 2016.
[69] Jonas Adler and Sebastian Lunz. Banach Wasserstein GAN. NeurIPS, 2018.
[70] Lucas Theis, Aäron van den Oord, and Matthias Bethge. A note on the evaluation of generative models. arXiv preprint arXiv:1511.01844, 2015.