論　文

機械学習を用いた離職予測に関する文献レビュー

三田寺 裕治

（受理日：2024年１月16日）

Literature Review on Turnover Prediction Using Machine Learning

Yuji MITADERA

要　旨

機械学習を用いた離職予測に関する先行研究をサーベイし、離職者を予測するためのアルゴリズムとその予測精度及び離職に影響を与える特徴量を明らかにした。アルゴリズム単体で集計すると、最も良く使われているアルゴリズムはRF（11/15：15文献中11文献で使用）であり、続いてLR、NB（9/15）、DT（8/15）、SVM、KNN（7/15）であった。アルゴリズムの要素や特徴で集計すると、最も良く使われていたのはアンサンブル学習であった（13/15）。データセットや前処理等が異なるため、単純に比較することはできないが、指標だけで比較すれば、最も精度が高かったアルゴリズムはRFであった（ACC=.9940, AUC=1.000）。

離職に影響を与える代表的な特徴量は給与・昇進・昇給、個人属性、内部要因（企業内要因）、有給取得、出張、福利厚生（寮）、残業であった。研究において使用されているデータセットのソースは極めて少数（１～２個）であり、一般化が不十分という問題点がある。特徴量については離職の大きな原因とされている職場の人間関係や社風・組織風土、ライフイベント、健康状態などがほとんど使用されていないことが問題点としてあげられる。

キーワード：機械学習、AI、離職予測、リテンションマネジメント

緒言

日銀短観の雇用人員判断D.I.から人手不足の状況を確認すると、2013年にマイナス値（不足超）となって以降、ほぼ一貫して下降傾向を示し2019年６月時点で▲32％ポイントとなっている[1)]。COVID-19の感染拡大により経済活動が停滞し2020年には一時的に人手不足感が弱まったが、５類感染症に変更されてからは経済活動が徐々に正常化に向かい、雇用人員判断D.I.は2023年６月時点で▲32％ポイントとなり[2)]、再び人手不足感が強まっている。令和４年版高齢社会白書によると、日本の生産年齢人口（15～64歳）は2050年には5,275万人（2021年から29.2％減）に減少すると見込まれており[3)]、今後も労働力不足が一層加速する可能性がある。

こうした状況下において企業が人材を確保するためには、採用活動を強化するだけでなく、既存社員の離職を防止し定着を図るための取り組みも重要となる。つまり、定期的に離職予測を行い、離職リスクの高い従業員に対して早期に介入や支援を行うことが求められる。離職予測を行う際には従業員の個別性を考慮する必要があるが、個別性には多くの要素があり、人間が手作業で予測するのは困難である。また、人事担当者が経験や勘に基づいて予測すると、予測精度にバラツキが生じる可能性がある。

そのため、近年では機械学習アルゴリズムを用いて離職リスクの高い従業員を予測・特定するなど、先進的なテクノロジーを用いてリテンションの問題を解決しようとする試みが始まっている。

しかしながら、こうした分野の研究は発展途上にあり、実用に耐えうる予測システムを構築する

には、離職予測に大きく影響する特徴の選択、精度が高く頑健な機械学習アルゴリズムの選択など、更なる研究の蓄積が求められる。

これらのことを踏まえ、本稿では機械学習を用いた離職予測に関する先行研究をサーベイし、離職者を予測するためのアルゴリズムとその予測精度及び離職に影響を与える特徴量を明らかにすることを目的とした。

方法

文献検索にはGoogle Scholarを使用し、“turn over” “prediction” “artificial intelligence” “machine learning” のキーワードを組み合わせてAND検索を行った。文献発行期間は2010年から2023年９月までとした。検索によって得られた論文及びハンドサーチによって抽出した論文について、抄録、本文を精読し、機械学習によって離職予測が行われ、その精度がACC（Accuracy：正解率）またはAUC（Area Under Curve：曲線下面積）によって評価されている15編の文献をレビュー対象とした。これら15編の論文について一覧表を作成し、発行年、筆者所属機関の所在国、予測時期、データセット、業種・職種、データサイズ、使用特徴／特徴量、特徴量数、特徴選択アルゴリズム、機械学習アルゴリズム、評価指標、bestACC、bestAUC、経営上の示唆、技術上の課題などについて整理・分析を行った。

結果

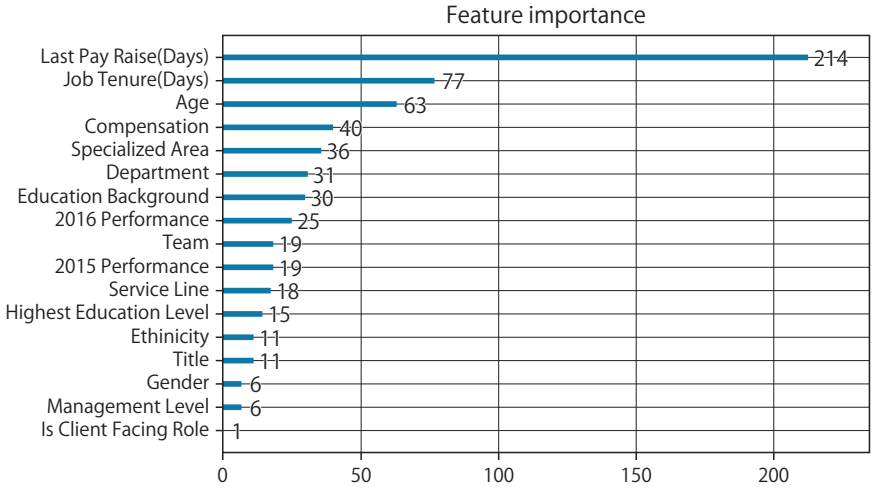
分析対象となった15編を発行年毎にみると、2015年１件、2016年１件、2018年２件、2019年２件、2020年１件、2021年５件、2022年２件、2023年１件であった（表１）。使用されたデータセットは米地銀の従業員情報、チリのコールセンター従業員の勤務記録、（Google Worspace等から収集した）従業員の電子メール・データセット、Oracleベースの ERP（統合基幹業務システム）従業員データセット、米グローバル小売企業の従業員情報、約100社の人事情報、１パッケージ製造企業の人事・離職・休暇取得データセット、ヨルダンの１サービス業企業の従業員を対象とした質問紙調査結果、韓国の１病院に勤務する看護師の人事情報、IBM HR analytics employee attrition dataset（Kaggle1）、HR Analytics dataset（Kaggle2）である。なお、Kaggleはデータサイエンスと機械学習のコンペティションプラットフォームで、数多くのデータセットが投稿されている。予測時期は在職中が14件、採用段階（求職中）が１件であった。各論文の概要は次の通りである。

Zhao[4)]らは、機械学習を用いた離職予測の先行研究をレビューし、①人事データは機密であり、複数のデータセットについての詳細な分析が困難である上、ノイズが多く、一貫性がなく欠損もあること、②多様なデータセットに対し、予測アルゴリズムの評価指標が偏っており、伝統的にACCが使われることが多いが、ACCは不均衡なデータセット[注1)]では信頼性が低いこと、③モデルの解釈（特徴の重要性の順位付けや分類ルールの可視化等）は、モデルによって手法が異なることを指摘した。これらの問題に対処するために、Zhaoらは、ノイズや欠損を含む多様なデータセットと複数のアルゴリズム、複数の評価指標を準備して実験を行い、信頼性の高いアルゴリズムの選択やモデルの解釈についての一般的なガイドラインを示すことにした。まず、米地方銀行従業員のデータセットとKaggle1を準備し、それぞれ前処理を行い、大中小のサイズにサンプリングして10個のデータセット[注2)]を作成した。そしてこれらに対する、10種類のアルゴリズム（DT, RF, GBT, XGB, LR, SVM, NN, LDA, NB, KNN）の予測結果を、５種類の指標（ACC, PRC, RCL, F1, AUC）により総合的に評価した。まず、各サイズごとにアルゴリズムのパフォーマンスを比較すると、小規模データセット（米地銀×２、Kaggle1×２）については、４個のデータセットに渡って、一貫して他よりも良いパフォーマンスが得られたアルゴリズムは存在しなかった。また中規模データセット（米地銀×２、Kaggle1×２）については、米地銀はXGB（AUC=.9634）、Kaggle1はNN（.7780）が最良であった。そして大規模データセット（米地銀×２）については、GBT（.9844～.9885）が最も好成績であった。次に、全データセットについて、アルゴリズムによってパフォーマンスに違いがあるかどうかを確かめるために、AUCの全データセットに渡る中央値についてKruskal-Wallis検定を行ったと

ころ有意であり（p=.002）、事後検定では、ツリー・ベースのアルゴリズム（XGB, GBT, RF, DT）、特にXGBとGBTが1、2番目に好成績であった（ .9462, .9417）。これらの結果を踏まえた実用的なガイドラインを述べると、まずサイズごとの最適なアルゴリズムについては、小規模データセットには特に適したアルゴリズムはなく、中・大規模データセットにはツリー・ベースのアンサンブル手法（XGB, GBT）が適している。また特徴量の重要度の順位付けに関しては、最も優秀なアルゴリズムを基にして、重要度を計算するのが良い。例えば、本研究で最も優秀だったXGBに基づいて、中規模米地銀を対象とした重要度を求めると、１、２、３番目に重要な特徴量は最終昇給からの日数、在職日数、年齢であった（図１）。最後に、分類ルールの可視化に関しては、ツリー・モデルが適している。ただしツリー・ベースと言ってもアンサンブル手法は、ツリーが複数存在するため可視化が困難であり、妥協案として一本のツリー選ぶべきである。

Korytkowskiら[5]は、１従業員の離職は他の従業員に伝染する可能性があることから、従業員の仕事上のコミュニケーションを分析することにより、離職予測が可能になると考えた。そして28か月にわたる仕事に関連した電子メール数千人分を収集し、28個の特徴量（パンデミック（新型コロナ流行中かどうか）、友人離職（メッセージを交換した相手の中で最近離職した者がいるかどうか）、各曜日の送信数、各曜日の受信数、同僚からの受信数、社外からの受信数、上司からの受信数、勤務時間外の受信数、職位[注3]からなるデータセットにまとめた。次に、特徴量選択機能をもつグリア・ネットワーク[注4]とそれと接続された畳み込みNNを用いて、重要度の高い特徴量を選択した。そして学習モデルとして２種類の回帰型NN（LSTMとGRU）[注5]を用意し、特徴量選択前のデータセットと、選択後のデータセットをそれぞれ学習させた結果を比較したところ、LSTM、GRUともに選択後の方がACCが高く（LSTM: .741→ .749, GRU: .732→.743）、学習時間も短かかった（LSTM: 26分→21分, GRU: 25分→22分）。この結果をまとめれば、従業員の電子メールを用いると、最大.74のACCで離職予測が可能であった。

ValleとRuz[6]は、離職率の高いチリのコールセンターの離職を抑えることを念頭に、保険販売代理店と携帯電話プラン販売代理店から、販売員のデータセット2,407人分を入手し、離職予測を行った。データセットには６つの特徴量（就業時間、通話時間、有効連絡数（有望な顧客との電話連絡。販売成功の指標）、販売数、完了記録数（販売に失敗し、電話連絡が完全に終了した数）、生産（販売金額、販売員へのインセンティブ、ボーナスの合計））が含まれる。またプロスペクト理論に基づいて、時間経過に伴う業績の変化が離職率に影響を与えると考え、販売数と生産の時間変化（販売数



出典：Yue Zhao, Maciej K. Hryniewicki, Francesca Cheng, et al. Employee Turnover Prediction with Machine Learning: A Reliable Approach. In Proceedings of SAI intelligent systems conference: 737–758, 2019.　より転載

図１　米地銀中規模データセットに対しXGBを適用した際の特徴重要度

表１　機械学習を用いた離職予測に関する文献の一覧

論文名	著者	筆者所属機関の所在国	発行年	予測時期	データセット	業種・職種	データサイズ（元データの従業員数→前処理後の従業員数）	特徴／特徴量数（特徴選択前→後）	使用特徴／特徴量（特徴削減前。下線：予測において重要度の高い特徴量）	特徴選択アルゴリズム	機械学習アルゴリズム	評価指標	best ACC（特徴選択後）	best AUC（特徴選択後）	経営上の示唆	技術上の課題
Employee Turnover Prediction with Machine Learning: A Reliable Approach	Yue Zhao, Maciej K. Hryniewicki, Francesca Cheng, Boyang, Xiaoyu Zhu	canada, germany, us	2019	在職中	1)米地銀の従業員情報, 2)Kaggle1 (IBM Watson Analytics simulated dataset)	1)銀行, 2)架空	1)14,322→9,089→小規模(50, 100), 中規模(500, 1000), 大規模(5000, 9000), 2)1,470→1,470→小規模(50, 100), 中規模(500, 1500)	1)24→19, 2)38→31	1)2015 Performance(2015年業績), 2016 Performance(2016年業績), LastPayRaise(最終昇給からの日数), JobTenure(在職日数), Age(年齢), Compensation(報酬), SpecializedArea(専門分野), Department(学部), EducationBackground(専攻), Performance(業績), Team(チーム), ServiceLine(業務), Highest EducationLevel(最終学歴), Ethnicity(エスニシティ), Title(役割), Gender(性別), ManagementLevel(職位), IsClientFacingRole(接客業務), 2)BusinessTravel(出張), DailyRate(日給率), Department(部署), DistanceFromHome(家からの距離), Education(教育), EducationField(専攻分野), EvnironmentSatisfaction(職場環境満足), Gender(性別), HourlyRate(時給率), JobInvolvement(仕事関与), JobLevel(職位), JobRole(職種), JobSatisfaction(仕事満足), MartialStatus(婚姻), MontlyIncome(月収), MontlyRate(月給率), NumCompaniesWorked(経験会社数), Over18(18歳超), OverTime(残業), PercentSalaryHike(給与上昇率), PerformanceRating(仕事評価), RelationshipSatisfaction(関係満足), StandardHours(標準労働時間), StockOptionLevel(ストックオプションレベル), TotalWorkingYears(総労働年数), TrainingTimesLastYear(昨年研修回数), WorkLifeBalance(ワークライフバランス), YearsAtCompany(勤続年数), YearsCurrentRole(在現職年数), YearsSinceLastPromotion(最終昇進からの年数), YearsWithCurrentManager(現上司との年数)	–	DT, RF, GBT, XGB, LR, SVM, NN, LDA, NB, KNN	ACC, PRC, RCL, F1, AUC	–	0.9462(XGB。全データセットの中央値)	–	特徴に関する処理。
Employee Turnover Prediction From Email Communication Analysis	Marcin Korytkowski, Jakub Nowak, Rafał Scherer, Anita Zbieg, Błażej Żak, Gabriela Relikowska & Paweł Mader	poland	2022	在職中	従業員の仕事関連email	–	数千	28→19～24（特徴量数）	Pancdmic(新型コロナ流行中か), IfAnyOfFriendsHasStoppedWorking(メッセージを交換した相手の中で最近離職した者がいるか), EmailsSent(各曜日の送信数), ReceivedEmails(各曜日の受信数), FromColleagues(同僚からの受信数), FromOutside(社外からの受信数), FromSupervisor(上司からの受信数), AfterWrokingHours(勤務時間外の受信数), PositionAtWork(職位)	Glial NN	NN	ACC, PRC, RCL	0.74	–	離職は他の従業員にも伝染する。重要人物の突然の離職は会社に大きなダメージを与えるため問題である。	–
Turnover Prediction in a Call Center: Behavioral Evidence of Loss Aversion using Random Forest and Naïve Bayes Algorithms	Maurício A. Valle & Gonzalo A. Ruz	chile	2015	在職中	チリコールセンターの勤務記録	コールセンター(保険販売、携帯電話プラン販売)	2,407	8	LoggedHours(就業時間), TalkedHours(通話時間), EffectiveContacts(有効連絡), NumOfApprovedSales(販売数), NumOfFinishedRecords(完了記録数), ApprovedProduction(生産), DifferenceBetweenApprovedSales(販売数差), DifferenceBetweenApprovedProduction(生産数差)（連続する２ヶ月間の差分）	–	RF, NB	RCL, PRC, ACC, AUC	.856(RF)	1.000(RF)	コールセンターは離職率が高く、新社員の募集、採用、訓練コストも高い。人員欠乏が常態化しており、従業員の過労も問題で、コストの低い海外への移転も激しい。本システムは離職予防の他、業績が悪い従業員の発見にも利用可能。	過去の業績データがなく、給与の期待値やスキルを測定する心理変数等から離職率を予測する。

論文名	著者	筆者所属機関の所在国	発行年	予測時期	データセット	業種・職種	データサイズ（元データの従業員数→前処理後の従業員数）	特徴／特徴量数（特徴選択前→後）	使用特徴／特徴量（特徴削減前。下線:予測において重要度の高い特徴量）	特徴選択アルゴリズム	機械学習アルゴリズム	評価指標	best ACC（特徴選択後）	best AUC（特徴選択後）	経営上の示唆	技術上の課題
Ensemble method based architecture using random forest importance to predict employee's turn over	Anwar Hossen, Emran Hossain, Zahereel Ishwar Abdul Khalib, Fatema Siddika	bangladesh, malaysia	2020	在職中	Kaggle2 (Kaggle HR Analytics dataset)	−	15,000	9→5	satisfaction_level(満足度), last_evaluation(直近評価), number_project(割り当てプロジェクト数), average_monthly_hours(平均月労働時間), time_spend_company(在社年数), work_accident(労災), promotion_last_5years(５年以内の昇進), salary, type(職種)	χ^2統計量, RF重要度	RF, DT, SVM, NN, GNB, KNN, Bagging, Boosting	ACC, AUC	.9940(RF)	.99(RF)	高い離職率は、技能をもつ従業員の離職、会社の評判の低下をもたらすため好ましくない。	−
Prediction of Employee Turn Over Using Random Forest Classifier with Intensive Optimized Pca Algorithm	Alaeldeen Bader Wild Ali	serbia	2021	在職中	Oracle ERP dataset	−	330	記載なし（7以上→4と推測される）	記載なし(Appointment date(就任日付), Appraisal ratings(評価), Basic details(基本情報), Competencies(能力), Exit reason(退職理由), Loan details(ローン情報), Qualifcation and Salary elements(資格・給与要素)と推測される)また、これらを主成分分析により４要因(内部要因、仕事満足度、外部要因、従業員状況)に集約した。	主成分分析	RF, LR, NB, KNN, DT	ACC, PRC, RCL, F1, AUC	.91(RF)	.916(RF)	−	人口動態変数間の関係の考慮、離職の心理的側面の研究。
A Machine Learning Approach to Analyze and Reduce Features to a Significant Number for Employee's Turn Over Prediction Model	Mirza Mohtashim Alam, Karishma Mohiuddin, Md. Kabirul Islam, Mehedi Hassan, Md. Arshad-Ul Hoque & Shaikh Muhammad Allayear	bangladesh, malaysia	2018	在職中	Kaggle2 (Kaggle HR Analytics dataset)	−	15,000	9→3	satisfaction_level(満足度), last_evaluation(直近評価), number_project(プロジェクト数), average_monthly_hours(平均月労働時間), time_spend_company(勤続年数), work_accident(労災), promotion_last_5years(5年以内の昇進), salary, type(職種)	逐次後方選択アルゴリズム(SBS), χ^2統計量, RF重要性	DT, RF, SVM, MLP, KNN, GNB	ACC, AUC	.9803(RF)	.99(RF)	退職に繋がる変数(満足度等)を改善することで、退職率を抑制できる。	雇用主との関係等も考慮する必要がある。
A Proposed Model for Predicting Employee Turnover of Information Technology Specialists Using Data Mining Techniques	Ahmed Ghazi, Samir Ismail Elsayed, Ayman Elsayed Khedr	egypt, saudi arab	2021	在職中	IBM HR Employee Attrition dataset (特徴量からKaggle1と推測される)	−	882	28以上	Age(年齢), DairyRate(規定日給), Department(部署), DistanceFromHome(家からの距離), Education(教育), EnvironmentSatisfaction(職場環境満足度), Gender(性別), HourlyRate(規定時給), JobInvolvement(勤務態度), JobLevel(職位), JobRole(職種), JobSatisfaction(仕事満足度), MartialStatus(婚姻), MonthlyIncome(月収), MontlyRate(規定月給), NumberOfCompaniesWorked(経験社数), OverTime(残業の有無), PercentSalaryHike(昇給率), PerformanceRating(業績評価), RelationshipSatisfaction(人間関係満足度), StocOptionLevel(ストックオプションレベル), TotalWorkingYear(総労働年数), TrainingTimeLastYear(昨年の研修回数), YearAtCompany(勤務年数), YearsInCurrentRole(在現職年数), YearsSinceLastPromotion(最終昇進からの年数), YearsWithCurrentManager(現上司との年数), WorkLifeBalance(ワークライフバランス)等	−	GLM, LR, GBT, FLM, DT, DL, NB, SVM, RF	ACC, PRC, RCL, F1	.879(GLM)	−	高い退職率は、従業員のオリエンテーション、訓練、開発、維持への投資回収の観点から望ましくない。	−

論文名	著者	筆者所属機関の所在国	発行年	予測時期	データセット	業種・職種	データサイズ（元データの従業員数→前処理後の従業員数）	特徴／特徴量数（特徴選択前→後）	使用特徴／特徴量（特徴削減前。下線:予測において重要度の高い特徴量）	特徴選択アルゴリズム	機械学習アルゴリズム	評価指標	best ACC（特徴選択後）	best AUC（特徴選択後）	経営上の示唆	技術上の課題
Early Prediction of Employee Turnover Using Machine Learning Algorithms	Markus Atef, Doaa S. Elzanfaly, Shimaa Ouf	egypt	2022	求職中	Kaggle1 (IBM HR analytics employee attrition dataset)	−	1,470	5	Salary(給与), Age(年齢), DistanceFromHome(自宅からの距離), MartialStatus(婚姻), Gender(性別)	−	KNN, RF	ACC, PRC, RCL, F1, SPE, FPR, AUC	.84(KNN)	.82(RF)	人事担当者は、求職者に本システムを活用し、意思決定の一助とできる。	−
Prediction of Employee Turnover in Organizations using Machine Learning Algorithms: A case for Extreme Gradient Boosting	Rohit Punnoose, Pankaj Ajit	india(データはアメリカ)	2016	在職中	米グローバル小売企業の従業員情報	小売店舗リーダーチーム	73,115データポイント（従業員数×在職4半期数）	33	Age(年齢), NumberOfPromotions(昇進回数), Pay(給与), PeerAttrition(同僚の減少), TimeSinceLastPromotion(最終昇進からの時間)等。	−	XGB, LR, NB, RF, SVM, LDA, KNN	AUC	−	.86(XGB)	分析結果を、従業員の流出防止のための施策に利用する。	離職防止施策の効果測定により、施策の有効性を検証、ディープラーニング等の応用、その実装可能性の研究。
A machine learning-based analytical framework for employee turnover prediction	Xinlei Wanga & Jianing Zhi	us	2021	在職中	1)Kaggle1 (IBM Analytics dataset), 2)Kaggle2 (Kaggle HR Analytics dataset)	−	1)1,470, 2)14,999,	1)34, 2)9	1)Age(年齢), BusinessTravel(出張), DailyRate(規定日給), Department(部署), DistanceFromHome(家からの距離), Education(教育水準), EducationField(専攻分野), EmployeeCount(従業員数), EmployeeNumber(従業員番号), EnvironmentSatisfaction(職場環境満足度), Gender(性別), HourlyRate(時給), JobInvolvement(勤務態度), JobLevel(職位), JobRole(職務), JobSatisfaction(仕事満足度), MaritalStatus(婚姻), MonthlyIncome(月収), MonthlyRate(規定月給), NumCompaniesWorked(経験企業数), Over18(18歳超), OverTime(残業), PercentSalaryHike(昇給率), PerformanceRating(業績評価), RelationshipSatisfaction(人間関係満足), StandardHours(標準労働時間), StockOptionLevel(ストックオプションレベル), TotalWorkingYears(総労働年数), TrainingTimesLastYear(昨年の研修回数), WorkLifeBalance(ワークライフバランス), YearsAtCompany(勤務年数), YearsInCurrentRole(在現職年数),YearsSinceLastPromotion(最終昇進からの年数), YearsWithCurrManager(現上司との年数), 2)satisfaction level(満足度), last evaluation(直近評価), number project(割り当てプロジェクト数), average montly hours(平均月労働時間), time spend company(勤務年数), work accident(労災), promotion last 5years(直近5年間の昇進), sales(売上), salary(給与)	−	DT, RF, Extra Trees, LightGB Machine, XGB, CatBoost, GBT, AB, KNN, QuadraticDiscriminantAnalysis, NB, LR, LDA, Ridge, SVM, Blending, Stacking	ACC, RCL, PRC, F1, AUC	1).9118 (stacker-top5-XGB), 2).9933 (stacker-top5-RF)	1).83 (stacker-top5-LDA), 2).99 (stacker-top5-RF)	高離職率は士気の低下、知識・経験の喪失、企業文化への信念の喪失等に繋がり、しかも容易に回復できない。	−
Employee Turnover Prediction Using Machine Learning	Lama Alaskar, Martin Crane & Mai Alduailij	saudi arabia, ireland	2019	在職中	Kaggle2 (Kaggle HR Analytics dataset)	−	14,999	9→6(SelectKBest), 8(RFE), 5(RF)	satisfaction level(満足度)、average monthly hours(平均月労働時間)、last evaluation(直近評価)、number projects(関与プロジェクト数)、promotion last 5 years(直近5年間の昇進)、time spent company(勤務年数)、department(部署)、work accident(労災)、salary(給与)：下線部は3つの特徴選択法が共通して選んだ特徴	SelectKBest, Recursive Feature Elimination (RFE), Random Forest (RF)	LR, DT, NB, SVM, AB	ACC, PRC, SEN, SPE, F1, AUC	.966(RF × SVM)	.96(SelectKBest × DT)	−	DLの分類、遺伝子アルゴリズムの特徴選択への活用。

論文名	著者	筆者所属機関の所在国	発行年	予測時期	データセット	業種・職種	データサイズ（元データの従業員数→前処理後の従業員数）	特徴／特徴量数（特徴選択前→後）	使用特徴／特徴量（特徴削減前。下線:予測において重要度の高い特徴量）	特徴選択アルゴリズム	機械学習アルゴリズム	評価指標	best ACC（特徴選択後）	best AUC（特徴選択後）	経営上の示唆	技術上の課題
Analysis and Prediction of Employee Turnover Characteristics based on Machine Learning	Heng Zhang, Lexi Xu, Xinzhou Cheng, Kun Chao, Xueqing Zhao	china, us	2018	在職中	約100社の人事データ（特徴量よりKaggle1と推測される）	−	1,100	28以上	Age(年齢), BusinessTravel(出張頻度), Department(部署), DistanceFromHome(家からの距離), Education(学歴), Education field(専攻), EmployNumber(従業員数), EnvironmentSatisfaction(職場環境満足度), Gender(性別), JobInvolvement(勤務評価), JobLevel(職位), Job Role(職種), JobSatisfaction(仕事満足度), MartialStatus(婚姻), MontlyIncome(月収), NumCompaniesWorked(経験社数), PercentSalaryHike(給与上昇率), Overtime(残業), PerformanceRating(業績評価), RelationSatisfaction(関係満足度), TotalWorkYears(総勤務年数), TrainingTimesLastYear(昨年の研修回数), WorkLife Balance(ワークライフバランス), Standard hours(標準労働時間), YearAtCompany(在社年数), YearsInCurrent Role(在職年数), YearsSinceLastPromotion(最終昇進からの年数), YearsWithCurrentManager(現上司との年数)	相関分析	LR, バギング(XGB, RF, SVM)	ACC	.872(LR), .8932 (bagging)	−	−	データの特徴を理解し、適切な前処理を行うことが重要。特徴に関する処理は（本研究では行われておらず）、将来の課題。
Employee Turnover Prediction: The impact of employee event features on interpretable machine learning methods	Thee Juvitayapun	thailand	2021	在職中	1企業の人事情報＋離職記録＋休暇記録	パッケージ製造会社	6,552	15以上	age(年齢), education(学歴), martial status(婚姻), service years(勤続年数), 業務, time in positions(在位期間), work days(労働日数), last year salary increase compares to market salary increase rate(市場に比した昨年の昇給率), ratio of annual leave hours on average(年次平均有給休暇時間比率)等	相関分析	LR, RF, GBT, XGB	ACC, PRC, RCL, F1, AUC	.9855(GBT)	.9953(GBT)	離職の企業への影響は、1)採用、研修、優秀な学生を集めるための奨学金費用、2)採用・研修中の効率低下、3)特定分野の専門家の再雇用の困難さによる競争力低下。	従業員イベントという動的情報を入れることで、予測の精度が高まる。
Turnover Prediction using Machine Learning: Empirical Study	Mohammad Masoud, Yousef Jaradat, Esraa Rababa and Ahmad Manasrah	jordan	2021	在職中	ヨルダンの１企業の従業員を対象とした質問紙調査結果	サービス部門	280	11	会社の人事管理プロセス、職場環境の質、キャリアの安定性・給与に関する質問	−	soft clustering algorithm	ACC	0.84	−	分析結果を利用して、人事部はどういう環境を改善すべきか認識し、離職率低下に努める必要がある。	−
Development of a Nurse Turnover Prediction Model in Korea Using Machine Learning	Seong-Kwang Kim, Eun-Joo Kim, Hye-Kyeong Kim,Sung-Sook Song, Bit-Na Park, Kyoung-Won Jo	korea	2023	在職中	韓国の１病院勤務看護師の人事情報	病院看護師	630→629(退職看護師)＋780→777(在職看護師)	8	Age(年齢), Distance(自宅職場間距離), Dormitory(寮の使用), Grade(職位), Income(収入・給与), Marry(婚姻), ResidentialArea(居住地域), Sex(性別), Team(部署)	−	DT, LR, RF	ACC, PRC, RCL, F1, AUC	.989(RF)	.97(RF)	看護師の高離職率は、看護ミス、患者死亡率、残った看護師の負荷、管理コスト等の増大、看護サービスの質の低下に繋がる。	看護師の性格、経験、組織文化、リーダーシップ等の個人的・組織的要因、経済状況や政策変更等の外的要因は考慮されていない。データも１病院からのみである。

差、生産差）を特徴量に含めた。例えば、販売数差はある月の販売数と翌月の販売数の差であり、これが負であれば、販売数は減少しており、離職する傾向が高いと考えられる。このデータセットに対して前処理を行い、1ヶ月分を用いてRFとNBに学習させ、10分割交差検証により2ヶ月目の離職率を予測し、4つの指標(ACC, PRC, RCL, AUC)により評価した。その結果、ACCを除く3指標はRFの方が高く（表2）、RFは少量（1ヶ月分）の学習セットでも良好な成果を得られることが示された。次に2ヶ月分を使って3ヶ月目の離職率の予測結果を評価したところ、RF、NBどちらも概ね上昇が見られた。特にNBは上昇が顕著であり(PRC: .511→.780, RCL: .558→.779)、十分な学習セットがあればNBも良好な成果が得られることが分かった。最後に販売数差と生産差も投入したところ、さらに指標は向上し、RF、NBともにACCは.8以上、AUCは.9以上の数値が得られた。特徴量の重要性については、RFにおいては生産差、販売数差、販売数、完了記録数などが重要であった。生産差、販売数差は負、販売数、完了記録数は多い方が、離職率は高かった。生産の差や販売数の差が負であれば、販売員の収入は減少するため、妥当な結果と言える。一方、販売数が多い方が離職率が高いのは、優秀な販売員は離職しやすいことを意味しており、引き留め策が必要なことを示唆している。この結果について、ValleとRuzは機械学習による離職予測は離職を回避するための従業員支援に役立つと述べている。また本研究では在職者の離職予測しか行っていないが、給与の期待値や業務以外の特徴量（この種の仕事の専門的なスキルを測る心理的変数等）を用いれば求職段階（採用段階）で候補者を選別するのにも役立つと述べている。

Hossenら[7]は、予測精度を上げるために、χ^2統計量とRF重要度を用いた特徴量削減を提案し、Kaggle2を対象に実験を行った。具体的には、まず1）データセットの前処理を行い、2）χ^2統計量とRF重要度に基づいて、それぞれ最も重要度の高い特徴量セットを抽出し、それらを統合して5つの特徴量（平均月労働時間、満足度、在社年数、過去5年間の昇進、割り当てプロジェクト数）を選び、3）6種類のアルゴリズム（DT、RF、SVM、NN、GNB、KNN）に入力し、10分割交差検証と、4）RF、DT、KNNを用いたアンサンブル学習（バギング、ブースティング）を行い、各段階直後でパフォーマンス評価を行った。その結果、全段階に渡って最も優秀だったのはRFであり、1）前処理後（9特徴量）ではACC=.9849、2）特徴量選択後（5特徴量）ではACC=.9864、3）交差検証後の段階ではACC=.9940であった。これは、主成分分析を用いて特徴量を削減した先行研究（.9803）よりも高い数値であった。また4）バギングについてはACC=.9800、ブースティングについてはACC=.9850であった。これらの結果について、Hossenらは、従業員の離職を決定する主な要因は労働時間、満足度、在社年数、過去5年間の昇進、割り当てプロジェクト数であること、特徴量を削減した後、むしろ精度は高くなっていること、本研究が提案した変数削減手法が既存手法よりも優れていることを指摘した。

Ali[8]は、特徴量を効果的に削減するためにIntensive Optimized PCA（集中最適化主成分分析）を利用することを提案し、ERP（統合基幹業

表2　RFとNBの予測精度比較

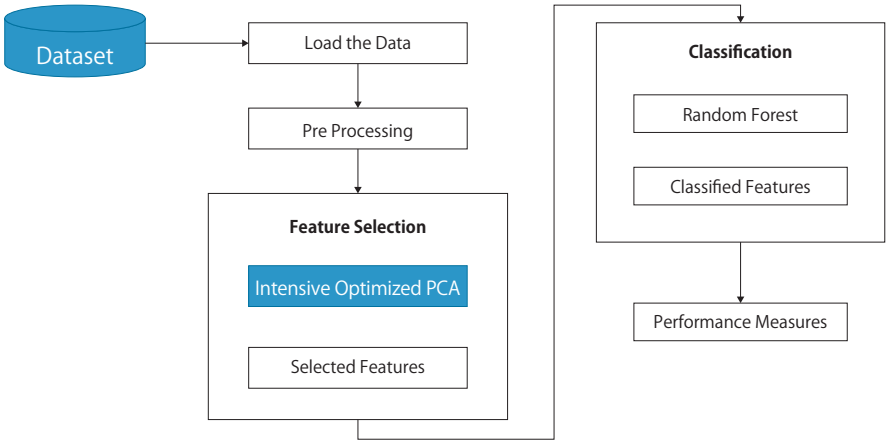
学習月数	使用特徴量	モデル	ACC	PRC	RCL	AUC
1	就業時間、通話時間、有効連絡数、販売数、完了記録数、生産	RF	.662	.816	.711	1.000
		NB	.664	.511	.558	.697
2	就業時間、通話時間、有効連絡数、販売数、完了記録数、生産	RF	.745	.630	.723	.999
		NB	.742	.780	.779	.812
2	就業時間、通話時間、有効連絡数、販売数、完了記録数、生産、販売数差、生産差	RF	.856	.813	.837	1.000
		NB	.828	.867	.833	.911

出典：Mauricio A. Valle, Gonzalo A. Ruz. Turnover Prediction in a Call Center: Behavioral Evidence of Loss Aversion using Random Forest and Naïve Bayes Algorithms. Applied Artificial Intelligence 29：923–942, 2015.　より筆者作成

務）システムから、従業員330人のデータセットを取得し、実験を行った（図２）。データセットは前処理された後、上記の主成分分析により、特徴が４要因（内部要因、仕事満足度、外部要因、従業員状況）[注6]に集約された。そして集約前のデータセットがLR、NB、KNN、DT、RF、集約後のデータセットがRFに入力され、パフォーマンスが比較された。その結果、評価指標のいずれも主成分分析を用いたRFが最も優秀であった（ACC=.91, PRC=.88, RCL=.99, F1=.93, AUC=.916）。またその要因により離職すると予測された従業員の数は、内部要因108人、仕事満足度42人、外部要因62人、従業員状況10人であり、離職予測には内部要因が重要なことが示された。これらの結果について、Aliは集中最適化主成分分析とRFを用いる本研究の手法は、高い精度で離職を予測することができ、しかも特定の業界やアプリケーションに拠らない利点があると主張した。

Alamら[9]は、データセットに含まれる変数を大幅に削減し、必要最小限の特徴量で機械学習予測を行うべきと考え、Kaggle2を実験対象に選んだ。９個の特徴量から離職と無関係と判断した労働災害を除いた上で、前処理を行い、SBS（Sequential Backward Selection）アルゴリズムを用いて特徴量を５個に減らし、さらにχ^2統計量とRF重要度に基づいて各特徴量の重要度を計算し、両者に共通する上位３つの３特徴量（満足度、平均月労働時間、在社年数）を選んだ上で、複数の機械学習アルゴリズム（DT、RF、SVM、NN、GNB、KNN）のパフォーマンスを比較した。このデータ処理過程における指標の変化を調べると、全過程を通して最も優秀なRFにおいては、特徴選択前（８特徴量）から特徴選択後（３特徴量）にかけて、ほとんど変化がなかった（ACC: 1.000→.9803, AUC: 1.00→.99）。これは他の大部分のアルゴリズムについても同様であり、２段階の特徴量削減が予測精度に与える影響は僅かであった。最後に３つの特徴量をそれぞれ軸とした３次元空間にデータセットをプロットして、離職群（離職すると予測された従業員のグループ）の特徴を可視化した。その結果RFでは、①全特徴量が低い群（労働時間は短いが、労働環境には不満な新人従業員）、②満足度が低く在社年数が短く、平均月労働時間が長い群（労働時間が長く、労働環境に不満な新人従業員）、③満足度が高く平均月労働時間が長く、在社年数は平均的な群（労働時間が長いが労働環境には満足している熟練従業員）が離職しやすいと予測された。Alamらは、最後の③の群が離職群と分類されたのは、熟練労働者は他社からの引き合いが多いからと推測している。そして満足度を高めるために報酬や研修を提供したり、新人従業員が新しい職場環境に適応できるよう、適切に扱う必要があると述べている。また研究の課題としては、従業員と雇用主の関係も特徴量としてモデルに含めるべきとしている。

Ghaziら[10]は、IBM HR Employee Attrition dataset[注7]を取り上げ、前処理を行った後、９種類



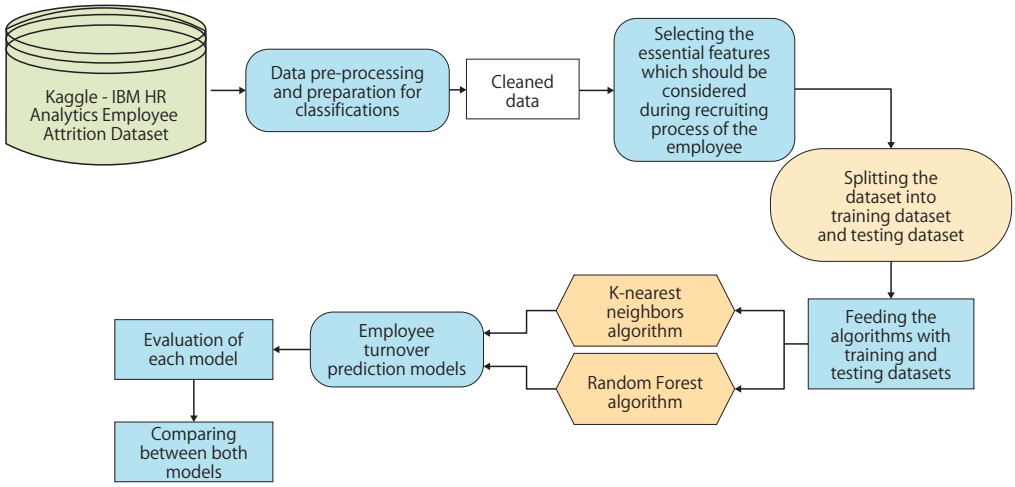
出典：Alaeldeen Bader Wild Ali. Prediction of Employee Turn Over Using Random Forest Classifier with Intensive Optimized Pca Algorithm. Wireless Personal Communications 119：3365–3382, 2021. より転載

図２　離職分類（予測）の流れ

の機械学習アルゴリズム（GLM、LR、GBT、FLM、DT、DL、NB、SVM、RF）に入力し、パフォーマンスを評価した。その結果、RCLはSVM、FLM、RFが最も優れていたが（RCL=1.000）、他の指標はGLMが最も優れていた（ACC=.879, PRC=.884, F1=.931）。次に各アルゴリズムについて特徴量の重要性を計算したところ、重要性の高い特徴量は年齢（DTで最高）、残業（GLM, DL, LR, NB, GBT, SVMで最高）、月収（FLM, RFで最高）等であり、年齢が低い、残業が多い、月収が低いほど、離職する傾向が強かった。

Atefら[11]は、求職者が採用後に離職するかどうかを予測するために、Kaggle1を用いて、複数の予測モデルの評価を行った（図３）。求職者の将来の離職に関連しうる特徴量（給与、年齢、自宅からの距離、婚姻、性別）を選択し、各特徴量と離職率の関係を調べたところ、年齢が低い、自宅からの距離が長い、給与が少ない、独身、男性の離職率が高い傾向が見られた。次にRFについて特徴量の重要度を計算したところ、離職率に大きな影響を与えていたのは給与、年齢、自宅からの距離であった。さらに相関を調べたところ、離職と比較的強い相関があったのは年齢（-.16）、給与（-.16）等であり、年齢や給与が低いほど離職率が高い傾向が見られた。最後にハイパーパラメータ（学習を制御するための設定変数。DTの分岐数やNNの層数等）を調整し、KNNとRFについて10分割交差検証を行い、パフォーマンスを評価した。その結果、KNNではACC=.84, PRC=.481, RCL=.126, F1=.196, AUC=.79、RFではACC=.80, PRC=.333, RCL=.249, F1=.281, AUC=.82であり、ACCとPRCにおいては、KNNの方が優れていた。また先行研究と比較すると、KNNにおいては、本研究がACC=.84, AUC=.79であるのに対し、先行研究ではACC=.832～.867, AUC=.52であり、本研究は先行研究と同等かより優れていた。またRFについては、本研究がACC=.80, AUC=.82であるのに対し、先行研究はACC=.85～.879、AUC=.58であり、本研究はACCでは少し劣っているが、AUCでは大幅に優れていた。Atefらは、この結果について、特徴量の数が少なく予測が困難であるにもかかわらず優れた結果が得られたのは、本研究のハイパーパラメータの調整方法が優秀だったからだと考察した[注8]。

PunnooseとAjit[12]は、ノイズが多い人事データに対応するために、頑健な機械学習アルゴリズムであるXGBの利用を提案した。そして米国グローバル小売企業から店舗リーダーチームの従業員データ（年齢、給与等）、米国労働統計局から統計データ（失業率、世帯収入の中央値等）を入手し、前処理を施し、７種類のアルゴリズム（XGB、LR、NB、RF、SVM、LDA、KNN）に入力し、AUC、学習時間、最大メモリ利用量を比較した。その結果、AUCとメモリ利用量が最も優秀であったのはXGBであった（AUC=.86、学習時間16分12秒、メモリ



出典：Markus Atef, Doaa Elzanfaly, Shimaa Ouf. Early Prediction of Employee Turnover Using Machine Learning Algorithms. International journal of electrical and computer engineering systems 13(2) : 135–144, 2022. より転載

図３　離職予測の流れ

利用量[注9)] 12%）。学習時間も大幅に長い訳でなく（7種類中4番目）、XGBは実際にノイズの多い人事データに対する離職予測に有効であることが示された。なお、PunnooseとAjitは離職予測だけでなく、離職予防施策についての研究の必要性も指摘している。

WangaとZhi[13)]は、データセットが単一で、学習モデルが少なく（1～10個）、特徴量エンジニアリング[注10)]も不十分な先行研究を批判し、データセットが複数で、学習モデルが多く、特徴量エンジニアリングも十分な研究を提唱した。そして特徴量エンジニアリング、モデルの学習と選択、アンサンブル学習を柱とする分析フレームワークに基づいて、Kaggle1とKaggle2を対象に実験を行った。まず特徴量について相関分析を行ったところ、離職と比較的強い相関があったのは、Kaggle1では残業（.25）、職位（-.17）、総勤務年数（-.17）Kaggle2では満足度（-.39）、給与（-.16）、労災（-.15）であった。またXGBに基づく重要な特徴量は、Kaggle1では月収、年齢、規定日給、Kaggle2では満足度、在社年数、プロジェクト数等であった。次にKaggle1に対して特徴量選択、Kaggle2に対しては特徴量エンコーディングと特徴量インタラクション（既存の特徴量を組み合わせて、新しい特徴量を作成すること）を行った上で、15個の基本モデル[注11)]についてF1を比較した。その結果、Kaggle1における上位5モデルは、LR（F1=.5140）、LDA（.4765）、ADA（.4568）、NB（.4365）、XGB（.4325）であり、Kaggle2では、RF（.9776）、ET（.9714）、LGBM（.9701）、XGB（.9693）、CB（.9668）であった。これら上位5モデルをブレンディングまたはスタッキングで組み合わせて、6個のアンサンブル学習モデル[注12)]を作成し、15個の基本モデルも交えてパフォーマンスを比較した。その結果、各指標で最も優秀だったのは、Kaggle1ではstacker-top5-XGB（ACC=.9118）、stacker-top5-NB（RCL=.7143）、stacker-top5-LR（PRC=.7619）、stacker-top5-LDA（F1=.6552）であり、Kaggle2ではstacker-top5-RF（ACC=.9933, PRC=.9919, F1=.9859）、stacker-top5-LGBM（RCL=.9720）であった。またF1で評価した最良モデル（Kaggle1はstacker-top5-LDA、Kaggle2はstacker-top5-RF）で計算すると、Kaggle1では離職者のうち在社と誤判断された割合は47.61%、在社者のうち離職と誤判断された割合は3.69%、Kaggle2ではそれぞれ2.98%、0.15%であった。さらに同じく最良モデルについてAUCを求めたところ、Kaggle1では.83、Kaggle2では.99であった。これらの結果をまとめれば、次のことが指摘できる。①基本モデルよりも、それらを組み合わせたアンサンブル学習モデルの方が優れている、②ブレンディング・モデルよりも、スタッキング・モデルの方が優れている、③Kaggle2よりもKaggle1の方が各指標が低い。その理由としてWangaとZhiは、学習データの少なさを挙げている（Kaggle1: 1,470, Kaggle2: 14,999）。

Alaskarら[14)]は、どのような特徴量選択法と機械学習アルゴリズムの組み合わせが、最もパフォーマンスが良いのか調べるために、Kaggle2を対象に実験を行った。まず特徴量の分布を調べたところ、割り当てられたプロジェクト数が少な過ぎる（2つ）または多過ぎる（6～7つ）場合や、営業・技術・サポート部門は離職率が高かった。次に生存分析を行ったところ、5年後の生存（在社）確率は、給与が低い従業員は約85%、中程度の従業員は約60%、高い従業員は約78%であった。また過去5年間に昇進した従業員は約90%、しなかった従業員は約50%であった。逆に言えば、中程度の給与や昇進しない場合、5年後の離職確率は高かった。続いて前処理を行い、SelectKBest[注13)]、RFE[注14)]、RFの3つの特徴量選択法によって、それぞれ重要な特徴量を求めたところ、SelectKBestでは満足度、平均月労働時間、労災、過去5年間の昇進、在社年数、RFEでは満足度、在社年数、直近評価、過去5年間の昇進、プロジェクト数、給与、所属（研究開発部門、人事部門、管理部門）、労災、RFでは在社年数、満足度、平均月労働時間、プロジェクト数、直近評価であった。次にハイパーパラメータを調整した後、これら3つの特徴量セットに対し、5つの機械学習アルゴリズム（LR、NB、SVM、DT、AB）を組み合わせた15ペアについて、10分割交差検証を行った。その結果、7つの指標（ACC、PRC、SEN、SPE、F1、AUC、誤分類）のうち、SENとAUCを除く5指標で最も良い成績を収めたのは、RF × SVM（ACC=.966, PRC=.944,

SEN=.913, SPE=.983, F1=.928, 誤分類=.3, AUC=.95)、SENとAUCで最も良い成績を収めたのはSelectKBest × DTであった（ACC=.948, PRC=.870, SEN=.918, SPE=.957, F1=.894, 誤分類=.5, AUC=.96)。これらの結果をまとめると、予測に重要な特徴は満足度や在社年数であり、予測精度が高いのはRF × SVMやSelectKBest × DTであった。

Zhangら[15]は、離職率を低減するのに有効な資料を提供することを目的として、離職率に影響を与える特徴を整理し、機械学習による予測実験を行った。まず約100社の離職率データ（使用特徴からKaggle1と思われる）に対し、まず相関分析を行って特徴量を削減し、前処理を施した後、LRモデルを用いて予測したところ、ACCは.872であった。またLRモデルの分析からは、離職に大きな影響を与える特徴は出張頻度、職種、専門分野、性別、婚姻、残業であり、頻繁に出張する従業員、販売員、人事や工学を専門とする従業員、男性、別居中の単身者、頻繁に残業をする従業員は離職確率が高かった。さらにXGB、RF、SVMによるアンサンブル学習（バギング）の結果、最終的にACCは.8932に向上した。

Juvitayapun[16]は、従業員のイベント（特に休暇）[注15]が離職に影響すると想定し、パッケージ製造会社従業員6,552人分の人事データ（年齢、性別、勤続年数、昇給等)、離職記録、休暇記録を収集し、年齢、勤続年数、在職期間、休暇記録、市場と比較した昇給率等の特徴量を持つデータセットに加工した。そして前処理[注16]を行い、４種類のアルゴリズム（LR、RF、GBT、XGB）に入力し、データの最終記録日から２ヶ月後の離職記録と照合し、離職予測を行い、休暇情報を使用する場合と、しない場合のパフォーマンスを比較した。その結果、休暇情報を使用しない場合に比べ、使用する場合は、LRはAUCが3.00%（.8674→.8934)、RFは4.48%（.8220→.8588)、GBTは13.14%（.8797→.9953)、XGBは10.46%（.8997→.9938）上昇し、ACC、PRC、RCLでも概ね上昇が見られた。さらにF1及びAUCに基づいてハイパーパラメータを調整したが、４種類のアルゴリズムとも、予測指標（ACC、RCL、PRC、AUC）にはほとんど改善は見られなかった。またXGBモデルを解釈すると、最も重要な特徴量は「市場昇給率と比べた昨年の昇給率」であり、離職群は在社群よりも小さかった。２番目と３番目に重要な変数は「離職１ヶ月前と離職２ヶ月前の年次有給休暇平均取得時間率（ratio of annual leave hours on average)」であり、離職群は在社群よりも、離職１ヶ月前の有給休暇平均取得時間率は1.5倍、離職２ヶ月前の有給休暇平均取得時間率は３倍高かった。これらの結果について、Juvitayapunは休暇イベントという特徴は予測精度の向上に有効であり、重要な特徴を特定することは、従業員の離職率を低減するための具体的施策を開発する際に有効であると述べている。

Masoudら[17]は、ヨルダンの１企業の離職率を研究するために、従業員280名を対象にアンケート調査[注17]を行い、回答に対してソフト・クラスタリング[注18]を適用し、従業員を３つのクラスタ（安定クラスタ、経験積みクラスタ、モバイル・クラスタ)[注19]に分類した。そして分類結果を調査した年の年末に得られた離職データと照合したところ、.84のACCが得られた。また回答分布の分析からは、従業員の多くは会社の人的資源管理の方法に強い不満を持っていること、公共部門や海外でのキャリア追求が離職の大きな要因になっていることが示された。この結果を受けてMasoudらは、企業の人事部に対し、金銭的インセンティブや従業員の努力や成果に対する経営陣からの感謝の表明を含む、様々な労働環境の改善を促した。

Kimら[18]は、看護師の離職に影響を与える要因を分析するために、韓国の三次病院の人事部門から、そこに勤務する看護師780名及び離職した看護師630名のデータ（年齢、性別、居住地域、寮の使用、婚姻状況、部門、雇用した年、離職した年、給与、雇用期間）を取得し、３種類のモデル（DT、LR、RF）を用いて離職予測を行った。その結果、ACC, PRC, RCL, F1で最も高いパフォーマンスを得たのはDT（ACC=PRC=RCL=F1=.92, AUC=.96)、ACC, PRC, F1, AUCで最も高いパフォーマンスを得たのはRF（ACC=PRC=F1=.92, RCL=.91, AUC=.97）であった。RFモデルはさらにハイパーパラメータを最適化され、最終的に.989のACCが得られた。またRFモデルにおいては、予測において最も重要度が高かった特徴は給与であり（.462)、次いで年

齢（.198）、寮利用（.196）であった。給与と年齢は低い方が、寮は利用している方が離職率は高かった。Kimらはこの結果について、給与が低いと看護師は自分の努力が認められなかったと受け取り、燃え尽きたりストレスを高めたりした結果、離職する可能性があり、公正な報酬を与える必要性を指摘した。また若い看護師は訓練や経験の不足、理論と実践の差等から、離職率を高めやすいため、十分な教育時間の提供や、理論と実践の差を埋める仕事の割り当て等を提案した。さらに寮の利用者は、家族親戚や知人からのソーシャルサポートを得られにくく、慣れない環境での生活の困難さから離職を選びやすいと考察した。

考察

1．使用されている機械学習アルゴリズム

アルゴリズム単体で集計すると、最も良く使われているアルゴリズムはRF（11/15：15文献中11文献で使用）、続いてLR、NB（9/15）、DT（8/15）、SVM、KNN（7/15）であった。アルゴリズムの要素や特徴で集計すると、最も良く使われていたのはアンサンブル学習であった（13/15）。この理由としては、比較的容易に高い精度が得られる等が考えられる。次に良く使われていたのは、DTをベースとしたアルゴリズム（DT、ET、RF、GBT、XGB、LightGBM、AdaBoost、CatBoost）であった（12/15）。その理由としては、木構造が予測結果の可視化に有利、前処理が比較的簡単等が挙げられる。その一方で、NNをベースとしたアルゴリズム（NN、DL）は使用文献が少なかった（4/15）。これは結果の解釈が難しい、精度が低い等が理由と考えられる。

2．パフォーマンス

データセットや前処理等が異なるため、単純に比較することはできないが、指標だけで比較すれば、最も精度が高かったアルゴリズムはRFであった（ACC=.9940[7]，AUC=1.000[6]）。これはRFのベースであるDTは、質的・量的双方を含む人事データに対応できるのが一因と考えられる。DTは単独では分類性能が必ずしも高くないが、RFはアンサンブル学習によってその欠点を補っていると言える。

3．代表的な特徴量

(1) 給与・昇進・昇給

離職に影響を与える代表的な特徴量についてみると、「給与」[11) 13) 18)]が離職に大きな影響を与えていることが明らかとなった。給与が低いと自分の能力や業績、貢献度が適切に評価されていないと考え、結果として離職しやすくなると考えられる。賃金と離職の関係については先行研究でも指摘されており、竹内らは給与が適切である（と従業員が認識している）ほど、従業員が所属施設を辞めて他の施設に移ろうという離転職意思が低下する[19)]と述べている。

市場昇給率と比べた昨年の昇給率も離職と関係していた[16)]。昇給とは在社年数や人事考課の結果などに基づいて従業員の賃金が増額することであるが、従業員の昇給率が市場昇給率よりも低い場合、従業員は現在の職場で自分の労働価値を十分に評価してもらえていないと認識し、転職する可能性が高まると考えられる。そのため、離職を防ぐためには、人事部門が業界内における給与水準（同エリア・同規模組織）を定期的に調査し、従業員の給与が業界の平均的な給与水準を下回らないように調整する必要がある。

LastPayRaise（最終昇給からの日数）[4)]や過去5年間の昇進[7)]も離職と関連があり、一定期間昇進や昇給がなく、今後も見込みがないと感じた場合、モチベーションが低下し離職する可能性が高まる。そのため、企業は定期的（最低1年に1度）に人事考課を行い従業員の勤務態度や能力、成果等を適切に評価する必要がある。昇進や昇給が一定期間ない従業員に対しては、その原因を特定するとともに、従業員の適性や興味に合わせて担当部署を変更したり、各従業員に合わせた個別の研修プログラムやメンタリングプログラムを提供し、従業員の職業的成長を支援していく必要がある。

(2) 個人属性

年齢によって離職の確率が異なることが示され、年齢が低い人の離職確率が高かった[4) 10) 11) 13) 18)]。これは日本におけるマクロデータでも同様の傾向が示されている。厚生労働省「令和3年雇用動向調査の概況」によると20～24歳の離職率は男性24.2%、女性26.9%、25～29歳では男性19.6%、女性17.9%

となっており、全年齢の平均離職率13.9％よりも高くなっている[20]。若年期は自分の適性や興味・関心に合った仕事を探したり、今後のキャリアの方向性を模索する時期であるため、離職率が高くなりやすい。若年層の早期離職を防ぐためには新しい環境に適応できるように組織全体でオンボーディングを実施することが重要である。また性別では男性[15]、婚姻では別居中の単身者[15]で離職確率が高かった。

経験については、Kimらが経験の浅い看護師は訓練や経験の不足、理論と実践の差から、離職率を高めやすい[18]と指摘している。看護師は大学や専門学校などの養成施設において看護師として必要な知識、技術を修得するが、実際の臨床現場では複雑かつ緊急性の高い状況に対応しなければならないことも多く、リアリティショックにより早期離職に繋がる場合がある。そのため、臨床経験の浅い新人看護師に対してはショックを緩和し臨床へのスムーズな移行を促す取り組みが求められる[21]。

(3) 内部要因（企業内要因）

「内部要因（企業内要因）」[8]も離職に影響を与えていた。企業は経営状態が悪化すると、余剰人員が多くなり、コスト削減の一環として従業員の解雇や契約の打ち切りなどが行われ、結果的に会社都合による非自発的離職が増加する。経営成績が悪化する理由は業界によって異なるが、経済不況や競合他社の出現による需要の減少といった外部環境の影響やロボティクスによる省力化、AIの進歩といった技術革新の影響が考えられる。

(4) 有給取得

休暇も離職に影響を与えていた。Juvitayapunによれば、従業員の「休暇（離職1ヶ月前と離職2ヶ月前の年次有給休暇平均取得時間率）」[16]が離職に影響していることが示され、離職群は在社群よりも、離職1ヶ月前の年次有給休暇平均取得時間率は1.5倍、離職2ヶ月前の年次有給休暇平均取得時間率は3倍高かった。これらの特徴を用いてリスクの高い従業員を特定し、早期に対応することで離職率を低減できる可能性がある。例えば、急に休暇が増えた従業員に対しては、その兆候を見逃さず面談や声掛けを行い、職務に関する不満を聞いたり改善策を一緒に考えることがあげられる。また、離職確率の高い優秀な従業員に対しては、給与の増額や重要ポストへの配置転換など、退職を引き止めるための提案・交渉（カウンターオファー）を行う必要がある。

しかし、離職1か月～2か月前は既に離職の意思が固まっており、手遅れである可能性も高い。そのため、普段から上司による面談を行ったり従業員の満足度やエンゲージメントを測定するなどして、従業員の満足度やモチベーションを把握しておくことも重要である。

(5) 出張

出張も従業員の離職率を高める要因となっていた[15]。出張は家族と過ごす時間を奪い、趣味の時間がとれなくなるなど従業員の生活満足度に影響を与える。また、出張では起床・就寝時間が不規則となりやすく、外食の増加や運動不足が健康に悪影響を及ぼす可能性がある。そうしたことが離職率に影響を与えているものと推測される。

出張が健康に及ぼす悪影響については先行研究[22]でも指摘されている。月に21泊以上出張した人は1～6泊の人と比べ肥満である確率が92％高く、拡張期血圧も高く、善玉コレステロールの数値が低かった。また、月に14泊以上出張した人は、月に1～6泊出張した人に比べ、BMIが有意に高く、健康に関する自己評価が低く、不安、抑うつ、アルコール依存症の臨床症状、運動不足、喫煙、睡眠障害などを報告する人が有意に高かった。企業は一人あたりの出張回数を減らしたり、リモートワークを積極的に推進するなど、出張の頻度を減らす対策を講じる必要がある。

(6) 福利厚生（寮）

看護師の離職を研究したKimは、寮を利用している看護師の方が離職率が高い[18]と指摘している。看護師寮は病院の近くにある場合が多いため通勤時間が短縮されたり、家賃が安いため住宅費を節約することができるなどメリットも多い。しかし、休日にも同じ職場の職員と顔を合わせることになり、仕事とプライベートの区別がつきにくくなる。また、友人など外部の者が入室できない寮の場合は職場以外の人との交流が希薄となり孤独感が高まる恐れがある。孤独感は抑うつや不安感といったメンタルヘルスの問題を引き起こし、離職率を

高める可能性があるため、看護師寮に居住している人は家族や友人など寮以外の人との交流を積極的に試みることが重要であると考えられる。

(7) 残業

残業の多い従業員は離職確率が高かった[10) 15)]。残業が多いと睡眠不足や慢性的疲労、メンタルヘルス不全など健康上の問題を引き起こす可能性がある。また、残業が多いとプライベートな時間を確保することができなくなり、ワークライフバランスを重視する従業員は不満が高まり離職を選択する可能性がある。企業は勤怠管理システムを導入し従業員の出勤、退勤時間を把握・管理することが重要である。また、仕事の量や難易度に応じて人的リソースを適切に配分し、過度な残業や労働負担を軽減していくことが求められる。

問題点及び今後の課題

まず、研究において使用されているデータセットについては、数が極めて少なく（１～２個）一般化が不十分という問題点がある。例えばRFは非常に高いパフォーマンスであったが、これはRFの構成要素であるDTには過学習という欠点があり、少数のデータセットに過剰適応し過ぎた結果、パフォーマンスが向上した可能性もある。そのため実用化に当たっては、多様なデータを用いて精度を検証する必要がある。また使用データセットは全て海外で得られたものであり、日本の実情に合っていない可能性もある。

予測時期については15文献中14件が在職中であり、採用段階（求職中）での離職予測は１件と少なかった。採用段階では予測に使用できるデータ（性別、年齢、学歴、職歴、保有資格、適性検査結果、面接試験結果など）が限られるため、精度の高い予測モデルを構築することが難しいが、採用段階で定着率の高い候補者や活躍する人材を予測できれば、採用コストや人材育成にかかるコストを大幅に削減することが可能となるため、さらなる研究の発展と深化が求められる。

特徴については離職の大きな原因とされている職場の人間関係（上司・同僚との相性など）や社風・組織風土、ライフイベント（結婚、出産、介護等）、健康状態、能力開発・資格取得の機会などがほとんど使用されていないのも問題点である。しかし、これらの情報の取得には困難が伴う。従業員間の人間関係や社風・組織風土を問う場合、従業員は今後のキャリアや人事考課への影響を懸念し真実を語らない可能性がある。ライフイベントや健康状態を問う場合も、プライバシー保護などの理由から、必ずしも必要なデータを取得できるとは限らない。また、離職予測に用いるデータが複数の部署（人事、人材開発、総務、財務・経理など）に分散して格納されている場合や、不適切に管理されている場合は、往々にしてデータ統合やクレンジングなどの処理が必要となる。

近年、必要最小限の仕事をして給料を得る「静かな離職」という現象がSNSで話題となっている[23)]。仕事に対するモチベーションが低く、最低限の仕事しか行わない従業員が増えれば、企業の競争力は低下し成長が阻害される。離職率は企業にとって重要なKPIであるが、企業が成長するためには離職率低減だけに着目するのではなく、優秀でパフォーマンスの高い従業員（ハイパフォーマー）やスタープレイヤーの離職をどのように防ぐかについて検討する必要がある。

文献

１）日本銀行調査統計局　短観（要旨）2019年６月 https://www.boj.or.jp/statistics/tk/yoshi/tk1906.htm（閲覧日：2023年10月）

２）日本銀行調査統計局　短観（要旨）2023年６月 https://www.boj.or.jp/statistics/tk/yoshi/tk2306.htm（閲覧日：2023年10月）

３）内閣府「令和４年版 高齢社会白書」p.4, 2022年

４）Yue Zhao, Maciej K. Hryniewicki, Francesca Cheng, et al. Employee Turnover Prediction with Machine Learning: A Reliable Approach. In Proceedings of SAI intelligent systems conference:737–758, 2019.

５）Marcin Korytkowski, Jakub Nowak, Rafał Schere, et al. Employee Turnover Prediction From Email Communication Analysis. Artificial Intelligence and Soft Computing:252–263, 2022.

６）Mauricio A. Valle, Gonzalo A. Ruz. Turnover

Prediction in a Call Center: Behavioral Evidence of Loss Aversion using Random Forest and Naïve Bayes Algorithms. Applied Artificial Intelligence 29:923–942, 2015.

7）Anwar Hossen, Emran Hossain, Abdul Khalib Zahereel Ishwar, et al. Ensemble method based architecture using random forest importance to predict employee's turn over. Journal of Physics: Conference Series 1755, 2020.

8）Alaeldeen Bader Wild Ali. Prediction of Employee Turn Over Using Random Forest Classifier with Intensive Optimized Pca Algorithm. Wireless Personal Communications 119:3365–3382, 2021.

9）Mirza Mohtashim Alam, Karishma Mohiuddin, Kabirul Islam, et al. A Machine Learning Approach to Analyze and Reduce Features to a Significant Number for Employee's Turn Over Prediction Model. Intelligent Computing: 142–159, 2018.

10）Ahmed Ghazi, Samir Ismail Elsayed, Ayman Elsayed Khedr. A Proposed Model for Predicting Employee Turnover of Information Technology Specialists Using Data Mining Techniques. International journal of electrical and computer engineering systems 12(2)：113–121, 2021.

11）Markus Atef, Doaa Elzanfaly, Shimaa Ouf. Early Prediction of Employee Turnover Using Machine Learning Algorithms. International journal of electrical and computer engineering systems 13(2)：135–144, 2022.

12）Rohit Punnoose, Pankaj Ajit. Prediction of Employee Turnover in Organizations using Machine Learning Algorithms:A case for Extreme Gradient Boosting. International Journal of Advanced Research in Artificial Intelligence 5(9)：22–26, 2016.

13）Xinlei Wanga, Jianing Zhi. A machine learning-based analytical framework for employee turnover prediction. Journal of Management Analytics 8(3)：351–370, 2021.

14）Lama Alaskar, Martin Crane, Mai Alduailij. Employee Turnover Prediction Using Machine Learning. Advances in Data Science, Cyber Security and IT Applications：301–316, 2019.

15）Heng Zhang, Lexi Xu, Xinzhou Cheng, et al. Analysis and Prediction of Employee Turnover Characteristics based on Machine Learning. The18th International Symposium on Communications and Information Technologies (ISCIT)：371–376, 2018.

16）Thee Juvitayapun. Employee Turnover Prediction: The impact of employee event features on interpretable machine learning methods. 13th International Conference on Knowledge and Smart Technology (KST)：181–185, 2021.

17）Mohammad Masoud, Yousef Jaradat, Esraa Rababa and Ahmad Manasrah. Turnover Prediction using Machine Learning: Empirical Study. International journal of advances in soft computing and its applications 13(1)：193–207, 2021.

18）Seong-Kwang Kim, Eun-Joo Kim, Hye-Kyeong Kim, et al. Development of a Nurse Turnover Prediction Model in Korea Using Machine Learning. Healthcare 11(11)：1583, 2023.

19）竹内規彦、竹内倫和、外島裕「人的資源管理研究へのマルチレベル分析の適用可能性：HRM 施策と組織風土が職務態度・行動に与える影響の検討事例」経営行動科学20(2)：127–141, 2007年

20）厚生労働省「令和３年雇用動向調査結果の概況」p.13, 2022年

21）谷口初美、山田美恵子、内藤知佐子、内海桃絵、任和子「大卒新人看護師のリアリティ・ショック―スムーズな移行を促す新たな教育方法の示唆―」日本看護研究学会雑誌 37(2)：71–79, 2014年

22）Andrew Rundle. Just How Bad Is Business Travel for Your Health? Here's the Data. Harvard Business Review. May 31, 2018. https://hbr.org/2018/05/just-how-bad-is-business-travel-for-your-health-heres-the-data

Andrew G Rundle, Tracey A Revenson, Michael Friedman. Business Travel and Behavioral and Mental Health. Journal of Occupational and Environmental Medicine 60(7):612–616, 2018. Catherine A Richards, Andrew G Rundle. Business travel and self-rated health, obesity, and cardiovascular disease risk factors. Journal of Occupational and Environmental Medicine 53(4):358–363, 2011.

23）齊藤豊「静かな離職と働かないおじさん」大妻女子大学人間関係学部紀要 24:1–15, 2022年

注

注１）離職者は在職者より大幅に少ないことが多いため、両者の比率は不均衡になる。

注２）米地銀データセットは従業員数14,322人、特徴数24個（全従業員で同じ値を取るものを削除して19個）、離職率28%であり、サンプリングして６個（小２、中２、大２）のデータセットを得た。Kaggle1は1,470人、38個（31個）、16%であり、サンプリングして４個（小２、中２）のデータセットを得た。

注３）送信数は７個、受信数は７個、職位は８個の特徴量を持つので、合計28特徴量である。

注４）グリア・ネットワークは、主NNの学習を制御する副次的なNNである。本研究では28個の特徴量を入力として受け付け、学習を行い、各特徴量の重要度を出力する。主NNはその情報を用いて、学習を効率化する。

注５）回帰型NNとは、ループを組み込んだNNであり、NNからの出力が同じNNへの次段階の入力となる。自然言語等の逐次型データの処理を得意とするが、情報を長期に渡って保持できないという欠点がある。LSTM（Long Short Term Memory）は、記憶セル、忘却ゲート、入力ゲート、出力ゲートをもつことにより、必要な情報を長期に保持できるようにした再帰型ネットワークである。GRU（Gated Recurrent Unit）は、LSTMから出力ゲートを省略する等、単純化した再帰型ネットワークである。両者は同様の性能を持つが、GRUの方が学習時間が短いとされる。

注６）内部要因（企業内要因）は余剰人員、配置転換、解雇、従業員の能力不足、契約終了等である。外部要因（企業外の社会的要因）は需要供給（人材の需給バランス、部品や製品等の需要・供給）、立地、環境、経済、政治等である。仕事満足度は給与評価（給与査定）、提案、仕事への不満等であり、従業員状況は健康状態ややむを得ない理由等である。

注７）使用されている特徴量から、Kaggle1と思われる。

注８）KNNでは通常、アルゴリズムそのものに対してハイパーパラメータを調整するのに対し、本研究では交差検証の度ごとにハイパーパラメータを調整すること、RFについては通常広範囲に渡って最適なパラメータを検索するのに対し、本研究ではパラメータの範囲を絞り込んでから検索することにより、好成績が得られたのだと推測した。

注９）メモリ総量は16GB。

注10）特徴エンジニアリングとは、特徴をエンコーディングしたり、インタラクションしたり、正規化したり、強調したり、選択したりすることである。特徴のエンコーディングとは、カテゴリー値をもつ特徴を、機械学習に適した変数に変換することである（例えば性別を、それぞれ0, 1の値をもつ男性を示す変数と女性を示す変数に変換する。性別=男→男性変数=1, 女性変数=0）。また変数のインタラクションとは、既存変数を組み合わせて新たな変数を作ることである。

注11）DT、RF、Extra Trees（ET）、Light Gradient Boosting Machine（LGBM）、XGB、CatBoost（CB）、GB、AB、KNN、Quadratic Discriminant Analysis（QDA）、NB、LR、LDA、Ridge、SVM

注12）blender-top5, stacker-top5-RF, stacker-top5-ET, stacker-top5-LGBM, stacker-top5-XGB, stacker-top5-CB。blender-top5は、上位５モデルの予測結果をブレンドして（組み合わせて）、最終的な予測結果を得るモデルであ

り、stacker-top5-RFは、RFを用いて５モデルをスタックし（積み重ねて）予測するモデルである。

注13）SelectKBestは、Pythonで実装された機械学習ライブラリーscikit-learnに含まれる変数選択メソッドで、分散分析に基づいている選択を行う。RFE(Recursive Feature Elimination)はロジスティック回帰、RFモデルはRF機械学習アルゴリズムに基づく選択手法である。

注14）Recursive Feature Elimination（再帰的特徴量除去）

注15）休暇に着目したのは、離職を考えている従業員は勤務態度が悪化したり、他社面接をしたりするため、休暇取得が増えると想定したからである。

注16）特徴削減を含む。特徴削減は、相互に相関の高い年齢、勤続年数、勤続年数に関連した特徴について、年齢を残して残りを削除した。

注17）質問内容は、非自発的離職（３項目)、企業の人事管理プロセス（４項目。「人事部門の人的資源管理の手法は適切である」）等)、労働環境（５項目。「労働環境は落ち着ける」等)、キャリアの安定性と給与（２項目。「公共部門でキャリアを積みたい」「海外で働く機会を得たい」）に関するものである。このうち非自発的離職以外の項目が、機械学習に使用された。５件法（５－強く同意する～1－まったく同意しない)。なお質問文は論文に掲載されておらず、論文記述からの推測である。

注18）ソフト・クラスタリングとは、データを分類する教師なし機械学習アルゴリズムである。通常のクラスタリングでは、クラスタ間の境界は明確であり、各データがどのクラスタに属するかは一意に決められるが、ソフト・クラスタリングでは明確でなく、各データの参加確率に応じて、そのデータがどのクラスタに属するか確率的に決められる。本研究では、各従業員の安定及びモバイル・クラスタへの所属確率を計算し、60%を超えたクラスタにその従業員を所属させる。60%を超えるクラスタがない場合は、経験積みクラスタに所属させる。

注19）安定クラスタは在社志向、経験積みクラスタは経験のための在社志向、モバイル・クラスタは転職志向の従業員の集団である。

Appendix 1　本研究で扱う機械学習アルゴリズムリスト

機械学習とはデータの背後にある予測や判断に利用できるパターンをコンピュータに自動的に発見させるAI技術である。そのためのアルゴリズムを機械学習アルゴリズムと言い、離職予測では次のものが良く使われる。

- Decision Tree（DT決定木)：単一の木構造を用いた機械学習アルゴリズム。学習結果（if thenルール）が木構造に沿って表現されるため、直観的に解釈しやすい。
- 線形モデル、線形分離をベースとした機械学習アルゴリズム
  - ➢ Logistic Regression（LRロジスティック回帰)：ロジスティック曲線を用いた回帰分析。データ（従業員など）が特定のクラスに所属する確率を出力する。実装が比較的簡単で、線形分離可能なデータの分類を得意とする。
  - ➢ Generalized Linear Model(GLM一般化線形モデル)：残差分布を自由に設定できる線形モデル（説明変数の線形結合で目的変数を説明する数理モデル)。線形回帰、ロジスティック回帰などが含まれる。
  - ➢ Support Vector Machine（SVMサポートベクターマシン)：データを２つのクラスに分離する超平面を創出することで、分類を行う手法。分類性能を高めるために、データが分離可能な別空間に特徴をマッピングする手法（カーネル法）が併用されることが多い。
  - ➢ Fast Large Margin（FLMファーストラージマージン)[1) 2)]：SVMを改良し、大量のデータや特徴に対応できるようにした手法。
  - ➢ Linear Discriminant Analysis(LDA線形判別分析)：低次元空間にマッピングすることで、データを異なるクラスに分離できる線形結合

を発見する手法。

- 神経細胞を模した機械学習アルゴリズム
  - ➢ Neural Network（NNニューラルネットワーク）：人間の神経細胞ネットワークの構造と動作をシミュレートするアルゴリズム。直列に接続された入力層、隠れ層、出力層を持ち、データを入力すると、そのデータが所属するクラスを出力する（層とは、神経細胞を模したユニットが複数並列に配置されたブロック）。多層パーセプトロン（Multi-Layer Perceptron: MLP）とも言う。
  - ➢ Deep Learning（DL深層学習）[3]：多数の隠れ層を持つNN。隠れ層は入力データの全体から細部までの多段階に渡る特徴を認識するのに使われ、高い分類精度をもつ。ただNNと同様、学習の成果はユニット全体に渡って保持されるため、その解釈は困難である。
- Naïve Bayes（NBナイーブベイズ）：ベイズの定理に基づいて、事前知識（従業員の特徴）から事後確率（ある特徴を持つ従業員が、あるクラスに所属する確率）を計算する手法。特徴の分布を表現するのに正規分布を用いる場合は、Gaussian Naïve Bayes（GNBガウシアン・ナイーブベイズ）と呼ばれる。
- K-Nearest Neighbors（KNN K近傍）：データがどのクラスに属するか、周囲のK個のデータの多数決によって決定する手法。
- アンサンブル学習アルゴリズム[4]：予測（分類）性能を高めるために、複数の学習モデルを組み合わせる手法。組み合わせ方には、バギングやブースティングなどがある。
  - ➢ バギング[4]：複数の学習モデルが並列して学習する手法。各モデルの出力を平均や多数決によりまとめることで、予測のばらつきを抑制できる。
  - ➢ ブースティング[5]：複数の学習モデルが直列（順番）に学習する手法。前のモデルのエラーを後続のモデルが修正することによって、予測の精度を高められる。
  - ➢ Random Forest（RFランダムフォレスト）[4]：学習モデルとしてDTを採用したバギング。DTよりも性能は良いものの、学習結果は複数の木構造で表現されるため、解釈しづらい。
  - ➢ Adaptive Boosting（ABアダプティブブースティング）[5]：前のモデルで誤って予測されたデータを特定し、後続のモデルではその重みを調整することで、学習エラーを最小化するブースティング。AdaBoost（アダブースト）とも言う。
  - ➢ Gradient Boosting（GB勾配ブースティング）[5]：勾配降下法を用いたブースティング。なおGBTは勾配ブースティング木のことである。
  - ➢ Extreme Gradient Boosting（XGB正則化勾配ブースティング）：正則化により過学習を抑えたGB。GBより高い精度と、短い計算時間を特徴とする。

Appendix 1 参考文献

1）RapidMiner, Inc. Fast Large Margin. Oct 31, 2023. https://docs.rapidminer.com/latest/studio/operators/modeling/predictive/support_vector_machines/fast_large_margin.html（閲覧日：2024年1月5日）

2）Rong-En Fan, Kai-Wei Chang, Cho-Jui Hsieh, et al. LIBLINEAR: A Library for Large Linear Classification. Journal of Machine Learning Research 9：1871–1874, 2008.

3）The MathWorks, Inc. ディープラーニング：これだけは知っておきたい3つのこと. https://jp.mathworks.com/discovery/deep-learning.html（閲覧日：2024年1月5日）

4）IBM. バギングとは. https://www.ibm.com/jp-ja/topics/bagging（閲覧日：2024年1月5日）

5）IBM. ブースティングとは. https://www.ibm.com/jp-ja/topics/boosting（閲覧日：2024年1月5日）

Appendix 2　評価指標

アルゴリズムやモデルの分類・予測性能を評価する指標には、次のものがある。

Accuracy（ACC正解率）：（TP+TN）/（TP+FP+TN+FN）。全予測の中で、正しく予測した割合。

Precision（PRC適合率）：TP / (TP+FP)。離職と予測した者のうち、実際に離職した割合。精度とも呼ばれるが、本稿では「精度」はACC、PRC等を含めた全般的な予測性能を指すこととする。

Recall（RCL再現率）：TP / (TP+FN)。実際の離職者のうち、正しく離職と予測した割合。

F-measure（F値, F1）：PRCとRCLの調和平均。双方が揃って高い場合のみ高くなるため、PRCとRCLを１つにまとめた指標と言える。

False Positive Rate（FPR偽陽性率）：FP / (FP+TN)。実際の在職者のうち、間違えて離職と予測した割合。

Area Under Curve（AUC）：横軸をFPR、縦軸をRCLとする二次元平面において、閾値（データ・ポイントを離職・在職を分類するための基準）を細かく変更した際得られる点（FPR, RCL）を連続的にプロットして作る曲線より下の面積。１に近いほど分類性能が高い（誤分類が少ない）

	P離職と予測	N在職と予測
P実際に離職	TP：True Positive（真陽性）	FN：False Negative（偽陰性）
N実際に在職	FP：False Positive（偽陽性）	TN：True Negative（真陰性）