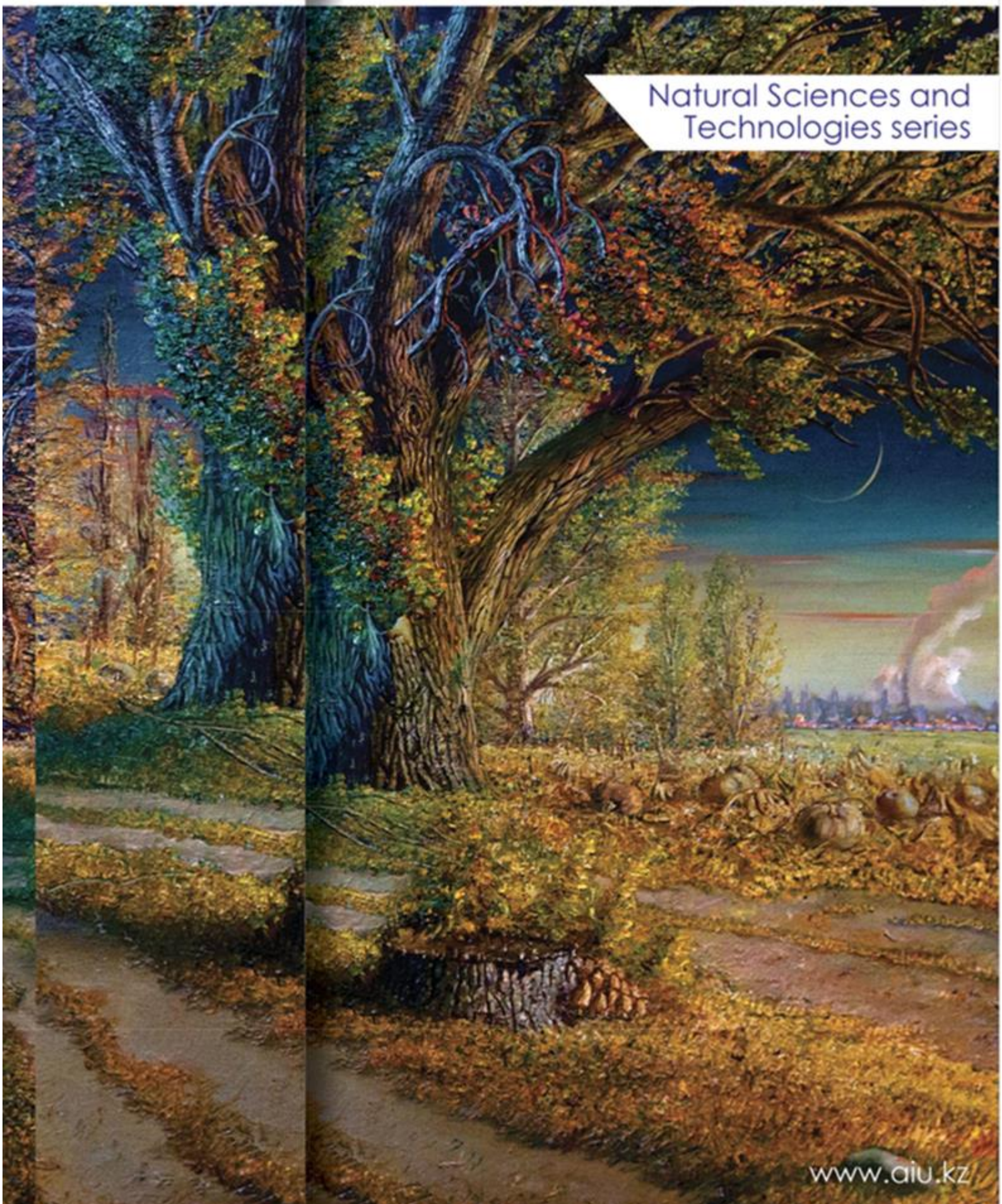
№4 (4) 2023

INTERNATIONAL SCIENCE REVIEWS

AIU

Natural Sciences and Technologies series



www.aiu.kz

ISSN: 2707-4862

ASTANA INTERNATIONAL UNIVERSITY

INTERNATIONAL SCIENCE REVIEWS
Natural Sciences and Technologies series

Has been published since 2020

№4 (4) 2023

Astana

Editorial address: 8, Kabanbay Batyr avenue, of.316, Nur-Sultan,
Kazakhstan, 010000
Tel.: (7172) 24-18-52 (ext. 316)
E-mail: natural-sciences@aiu.kz

International Sciense Reviews NST - 76153
International Science Reviews
Natural Sciences and Technologies series
Owner: Astana International University
Periodicity: quarterly
Circulation: 500 copies

CONTENT

СРАВНИТЕЛЬНЫЙ АНАЛИЗ МЕТОДОВ КЛАССИФИКАЦИИ В ИНТЕЛЛЕКТУАЛЬНОМ АНАЛИЗЕ ДАННЫХ

Р. Смадинов

Международный университет Астана, Астана, Казахстан
E-mail: r.smadinov@gmail.com

Аннотация. Данное исследование фокусируется на сравнительном анализе методов классификации в интеллектуальном анализе данных, используя данные, связанные с прямым маркетингом проводимые португальским банковскими учреждениями. Набор данных охватывает разнообразные аспекты клиентов, от демографических параметров до финансовых показателей. Основная цель заключается в извлечении информации из исторических данных для разработки прогностической модели и сравнительном анализе пяти моделей классификации: логистической регрессии, случайного леса, дерева решений, классификации SVC и k ближайших соседей. Результаты показывают, что случайный лес достигает наилучшей точности подчеркивая превосходство ансамблевого подхода, в то время как другие модели, такие как логистическая регрессия и k ближайших соседей, также проявляют высокую производительность в данном контексте.

Ключевые слова: интеллектуальном анализе данных, классификация, логистической регрессии, случайный лес, дерево решений, k ближайших соседей, классификации SVC

ВВЕДЕНИЕ

В современном информационном обществе, научные исследования в области интеллектуального анализа данных занимают центральное место, играя ключевую роль в принятии обоснованных решений в различных областях, начиная от бизнеса и финансов и заканчивая медицинскими науками. Одним из важных направлений в этой области является прогнозирование, и в частности, задачи классификации, которая направлена на правильное прогнозировании классов объектов. Существует множество методов классификации, предложенных и разработанных исследователями, каждый из которых обладает своими преимуществами и ограничениями. Следовательно, сравнительный анализ этих методов становится неотъемлемой частью процесса выбора оптимальной модели для конкретной задачи. Научная обоснованность и систематизация данного процесса могут значительно улучшить качество прогнозов и обеспечить успешное решение реальных проблем. В центре этого исследования лежит набор данных, полученный в результате деятельность прямого маркетинга, проведенных известным португальским банковским учреждением.

Рассматриваемый набор данных имеет большое количество элементов, обеспечивающих детальное понимание атрибутов клиента и результатов

кампании. Данные содержат как основные демографические параметры, такие как возраст, профессиональная деятельность, семейное положение и уровень образования, так и ключевые финансовые показатели, вроде баланса счета, статуса ипотечного кредита и истории невозврата кредитов. Помимо этих фундаментальных элементов, набор данных включает в себя такие нюансы, как характер общения, временную динамику, например, день последнего контакта, продолжительность взаимодействия. Этот разнообразный набор переменных закладывает основу для построения прогнозной модели, направленной на выявление закономерностей в поведении клиентов во время кампаний прямого маркетинга, уделяя особое внимание определению их вероятности подписки на срочные депозиты.

Основная цель данного исследования состоит в двух аспектах. Во-первых, оно направлено на извлечение информации из исторических данных, выявление тенденций, лежащих в основе успешных и неуспешных результатов подписки. Эти знания являются основой для разработки надежной прогностической модели, способной предвидеть будущее поведение подписчиков. Во-вторых, в проведении сравнительного анализа различных методов классификации, выявление их преимуществ и недостатков, а также предоставление исследователям и практикам базового понимания при выборе, наиболее подходящего моделей для конкретной задачи.

МАТЕРИАЛЫ И МЕТОДЫ ИССЛЕДОВАНИЯ

Исследование направлено на сравнительный анализ различных методов классификации в контексте интеллектуального анализа данных, используя данные, связанные с прямым маркетингом проводимые португальским банковскими учреждениями. Набор данных включает в себя разнообразные характеристики, каждая из которых служит важнейшим элементом для понимания поведения клиентов во время этих кампаний. В ходе исследования будет осуществлено построение и анализ пяти моделей классификации, включая логистическую регрессию, алгоритм случайного леса, дерево решений, классификация SVC и k ближайших соседей. Эти модели были выбраны в силу своей широкой применимости в задачах классификации.

Для построения и обучения моделей использованы библиотеки и инструменты машинного обучения, такие как scikit-learn в языке программирования Python. Обработка данных, включая преобразование и нормализацию, проведена с использованием стандартных методов предобработки данных. Оценка качества моделей проведена с использованием метрики точности (Accuracy), а также с помощью метода кросс-валидация (Cross-Validation). Метрики точность предоставляет из себя основной и простой критерий, используемый для измерения производительности модели интеллектуального анализа данных.

ДАННЫЕ

Данные были взяты с популярного источника информации Kaggle. Набор данных, используемый в данном исследовании, предлагает подробный и многогранный взгляд на взаимодействие с клиентами во время маркетинговых кампаний. Богатство набора данных заключается в разнообразии параметров, отражающих различные аспекты как демографических, так и поведенческих характеристик. Ниже представлен подробный информация об основных параметров:

Демографические параметры:

- Возраст (Age): Отражает возраст клиентов.
- Работа (Job): Описывает род занятий клиента, предоставляя информацию об экономических ролях, которые играют люди.
- Семейное положение (Marital Status): Описывает семейное положение клиентов.
- Образование (Education): Указывает на образование клиентов

Финансовые параметры:

- Дефолт (Default): Бинарный индикатор, показывающий, есть ли у клиента дефолтный кредит, отражающий его кредитоспособность.
- Баланс (Balance): Количественная оценка остатка на счете клиента - ключевой финансовый показатель, влияющий на принятие решений о подписке.
- Жилищный кредит (Housing Loan): Бинарный параметр, сигнализирующий о наличии у клиента действующего жилищного кредита и дающий представление о его финансовых обязательствах.

Коммуникационные данные:

Тип связи (Contact Communication Type): Перечисление методов, используемых для связи с клиентом (например, телефон, сотовая связь), с указанием каналов взаимодействия.

Временные параметры:

- День (Day): Указывается день месяца, когда произошел последний контакт с клиентом.

- Продолжительность (Duration): Количественная оценка продолжительности (в секундах) последнего контакта с клиентом во время кампании, отражающая уровень вовлеченности.

Параметры, связанные с кампанией:

- Счетчик контактов кампании (Campaign Contacts Count): Перечисление количества контактов, сделанных во время конкретной кампании для каждого клиента, что указывает на интенсивность работы с ним.

- pdays: Обозначает количество дней, прошедших с момента последнего контакта с клиентом в рамках предыдущей кампании, отражая временную динамику.

- poutcome: Описывает результат предыдущей маркетинговой кампании, обеспечивая исторический контекст для прогностического моделирования.

Всеобъемлющий характер набора данных делает его ценным ресурсом для обучения прогностических моделей и проведения тонкого сравнительного анализа методов классификации.

Таблица 1. Первичный вид данных

age	**job**	**marital**	**education**	**default**	**balance**	**housing**	**loan**	**contact**
30	unemployed	married	primary	no	1787	no	no	cellular

day	**month**	**duration**	**campaign**	**pdays**	**previous**	**poutcome**
19	oct	79	1	-1	0	unknown

МОДЕЛИ

Логистическая регрессия (Logistic Regression) - надежный выбор для сценариев, требующих простоты и интерпретируемости. Сильная сторона модели заключается в прозрачном представлении влияния признаков через коэффициенты, что делает ее особенно полезной для донесения информации до заинтересованных сторон [1]. Модель эффективна с вычислительной точки зрения, что очень важно для работы с большими массивами данных. Однако логистическая регрессия предполагает линейную границу принятия решения, что может

ограничить ее способность отражать сложные взаимосвязи в комплексных данных. Кроме того, ее производительность может быть чувствительна к выбросам, а предположение о независимости признаков может повлиять на точность в случаях, когда эти предположения нарушаются.

Классификатор случайного леса (Random Forest) – это хорошая модель при поиске мощного и универсального метода ансамблевого обучения [2]. Его сильная сторона заключается в том, что он уменьшает чрезмерную подгонку, объединяя прогнозы нескольких деревьев решений. Кроме того, он дает представление о важности признаков, помогая выявить критические факторы, влияющие на эффективность прогнозирования. Модель адаптируется к задачам классификации и регрессии и отлично справляется со сложными наборами данных. Однако ее интерпретируемость затруднена сложностью, которую вносят многочисленные деревья, а вычислительная трудоемкость обучения многочисленных деревьев может стать недостатком для больших наборов данных.

Классификаторы дерева решений (Decision Tree) отличная модель в связи со своей простотой и присущей интерпретируемость [3]. Эта модель эффективно отражают нелинейные взаимосвязи в данных и может работать с нерелевантными признаками без существенного снижения производительности. Прозрачность в принятии решений делает их подходящими для сценариев, в которых понимание обоснования модели имеет решающее значение. Тем не менее, деревья решений склонны к чрезмерной подгонке, улавливая шум в обучающих данных. Их нестабильность, проявляющаяся в изменении структуры дерева при незначительных изменениях данных, создает проблему для поддержания согласованности.

Классификатор опорных векторов (SVC) демонстрирует эффективность в высоко размерных пространствах и особенно эффективен при работе со сложными наборами данных [4]. Его сила заключается в адаптивности к различным функциям ядра, что позволяет моделировать сложные взаимосвязи в данных. Кроме того, SVC менее чувствителен к выбросам, что способствует его устойчивости в условиях зашумленных данных. Однако его вычислительные затраты могут быть ограничением, особенно при работе с большими наборами данных, и он может проявлять чувствительность к шуму, что сказывается на общей производительности.

Классификатор k ближайших соседей (KNeighbors) отличается простотой и интуитивностью, что делает его доступным выбором для различных приложений [3]. Его непараметрическая природа позволяет ему хорошо адаптироваться к локальным закономерностям и вариациям в данных. Однако ее вычислительные затраты растут с увеличением размера обучающего набора данных, что потенциально ограничивает ее масштабируемость. Кроме того, модель может быть

чувствительна к нерелевантным или избыточным признакам, что влияет на ее производительность в сценариях, где выбор признаков имеет решающее значение.

МЕТРИКИ

Точность (Accuracy) - это широко используемая метрика, которая обеспечивает прямое измерение общей корректности предсказаний модели. Ее сила заключается в простоте и легкости интерпретации, что делает ее основной метрикой для многих задач классификации. Она напрямую определяет отношение количества правильных предсказаний к общему количеству предсказаний, что дает четкое представление о производительности модели [5]. Однако точность имеет свои ограничения, особенно при работе с несбалансированными наборами данных. В сценариях, когда один класс значительно превосходит другие, достижение высокой точности может ввести в заблуждение, поскольку модель может быть предвзята к классу большинства. Поэтому точность сама по себе может быть не самой подходящей метрикой для оценки моделей в ситуациях, когда распределение классов неравномерно.

Кросс-валидация (Cross-Validation) - это мощный метод оценки обобщаемости модели путем разбиения набора данных на несколько подмножеств [6]. Ее сила заключается в том, что она дает более надежную оценку эффективности модели, чем одно разбиение на тренировки и тесты. Благодаря итеративному обучению и тестированию модели на разных подмножествах кросс-валидация помогает смягчить влияние изменчивости и случайности набора данных. Она особенно полезна в ситуациях, когда набор данных ограничен, что повышает надежность показателей эффективности.

РЕЗУЛЬТАТЫ

Для начала все количественные данные были рассмотрены с точки зрения описательной статистики в целях выявления каких выбросов в данных. Эта важно так как в случаи присутствие этих выбросов наши модели могут показать неверные результаты.

Таблица 2. Описательная статистка

	age	**balance**	**day**	**duration**	**campaign**	**pdays**	**previous**
count	49 732	49 732	49 732	49 732	49 732	49 732	49 732
mean	40,96	1 367,76	15,82	258,69	2,77	40,16	0,58
std	10,62	3 041,61	8,32	257,74	3,10	100,13	2,25

min	18,00	-8 019,00	1,00	0,00	1,00	-1,00	0,00
0,25	33,00	72,00	8,00	103,00	1,00	-1,00	0,00
0,5	39,00	448,00	16,00	180,00	2,00	-1,00	0,00
0,75	48,00	1 431,00	21,00	320,00	3,00	-1,00	0,00
max	95,00	102 127,00	31,00	4 918,00	63,00	871,00	275,00

Как видно из таблицы 2 наблюдается явные выбросы по колонкам «balance», «duration». Для решения проблемы можно в колонке «balance» удолить значение которое сильно отклонятся, а для колонки «duration» значения можно перевести значения из секунд в минуты.

Далее для работы с моделями необходимо категориальные данные перевести в бинарные значения, то есть в значение 0 и 1. Для этого значения по колонке «education» имеющие значения «secondary», «tertiary» были поставлены как 1, а остальные значения были поставлены как 0. Женатым было присвоено значение 1, а не женатым 0. Колонка «job» была разделена на несколько. Всем, у кого работа была указан как «management» была создана новая колонка под названием «manager» где менеджерам было присвоенное значение 1, а остальным 0, это было связано с тем, что в наборе данных чаше всего встречали люди работающие «management». Также была создана колонка «self_employed» где значение 1 присваивалось всем тем, у кого в колонке «job» было значение «employed» либо «entrepreneur», а остальным 0. Еще была создана колонка «blue-collar» где 1 ставилось всем кто в колонке «job» имел значение «blue-collar», а остальным был присвоен 0. Вместе с тем значение по колонке «poutcome» было разделено на 2 новые колонке. Первая колонка «cam_success» куда по значение 1 ставилось рядам, где значение по «poutcome» было «success», а остальным 0. Вторая колонка «cam_fail» куда по значение 1 ставилось рядам, где значение по «poutcome» было «failure», а остальным 0. По колонке «housing» значение 1 было дано всем, у кого есть жилищный кредит, а остальным соответственно 0. Колонка «default» была поделена таким образом что 1 ставилось всем, у кого дефолт и 0 тем, у кого его нету.

Таблица 3 – После преобразования

age	**default**	**campaign**	**abv_primary**	**married**	**manager**	**self_employed**
30	0	1	0	1	0	0

blue_collar	cam_success	cam_fail	has_loan	over_two_months	bal_over_4000	deposit
0	0	0	0	0	0	0

После этого к данным были применены причисленные модели после чего были рассчитана метрика точность (Accuracy) и далее эти данные кросс-валидированы в результате вышли следующие результаты.

Таблица 4 – Результаты по моделям

Model	Folder_1	Folder_2	Folder_3	Folder_4	Folder_5	MEAN
Logistic_Regression	0,8975	0,9038	0,9032	0,9028	0,8947	0,9004
Random_Forest_Classifier	0,8983	0,8999	0,9039	0,9065	0,8976	0,9012
Decision_Tree_Classifier	0,8657	0,8674	0,8631	0,8703	0,8703	0,8674
SVC	0,8850	0,8870	0,8864	0,8862	0,8847	0,8858
KNeighbors_Classifier	0,8926	0,8911	0,8893	0,8943	0,8907	0,8916

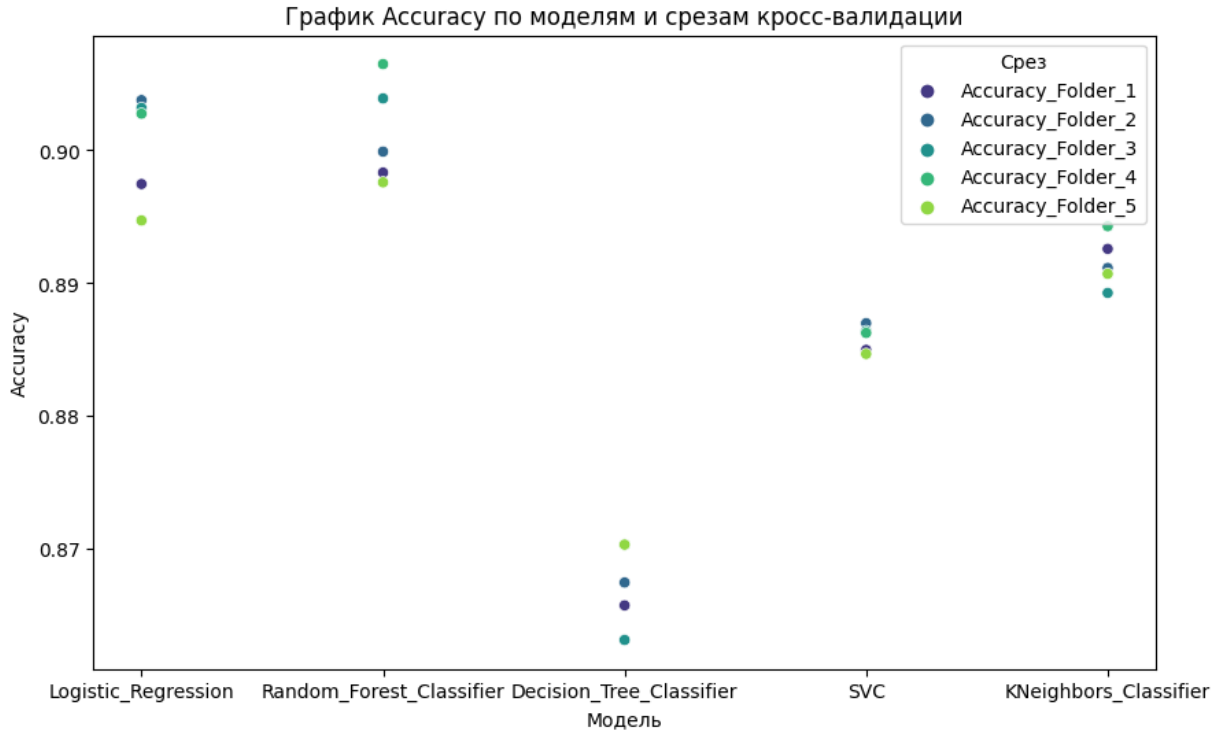


Рисунок 1. Результат по моделям

В результате классификационных моделей, созданных на этом наборе данных, классификатор случайного леса (Random Forest) показал лучшую точность, достигнув 90,12%. Его превосходство заключается в ансамблевом подходе к обучению, который использует коллективную силу нескольких деревьев решений, в результате чего получается надежная и точная прогностическая модель. Вслед за ним логистическая регрессия (Logistic Regression) продемонстрировала достойную производительность с точностью 90,04%, что подчеркивает ее надежность и простоту в отражении линейных связей в данных. Классификатор k ближайших соседей (KNeighbors), хотя и отстал от лидеров, продемонстрировал неплохую эффективность с точностью 89,16%. Его сила заключается в адаптации к локальным закономерностям, что делает его устойчивым выбором для улавливания тонкостей в наборе данных. Классификатор опорных векторов (SVC) и классификатор дерева решений (Decision Tree), достигли точности 88,58 % и 86,74 %, соответственно. SVC продемонстрировал эффективность в высокоразмерных пространствах, в то время как классификатор дерева решений продемонстрировал интерпретируемость, но с несколько меньшей точностью.

ВЫВОДЫ

При подведении итогов данного анализа данных становится очевидным, что различные модели обладают разными преимуществами. Этот сравнительный анализ подчеркивает важность выбора моделей, адаптированных к нюансам набора данных. Несмотря на то, что Random Forest и Logistic Regression лидируют, у каждой модели есть свои сильные и слабые стороны. Выбор между ними зависит от таких факторов, как интерпретируемость, эффективность вычислений и сложность взаимосвязей в данных. При навигации по ландшафту интеллектуального анализа данных выводы, сделанные на основе этих моделей, служат ориентирами для компаний, стремящихся оптимизировать маркетинговые усилия. Понимая прогностические возможности каждой модели, организации могут стратегически адаптировать свои подходы, сосредоточившись на методах, которые соответствуют характеристикам, присущим их данным. В завершение анализа становится очевидным, что классификатор Random Forest является маяком точности прогнозирования. Однако к выбору "лучшей" модели следует подходить взвешенно, учитывая конкретные требования и контекст решаемой задачи.

СПИСОК ЛИТЕРАТУРЫ

1. Park, H.-A. An introduction to logistic regression: from basic concepts to interpretation with particular attention to nursing domain. Journal of Korean Academy of Nursing, vol. 43(2), pp. 154–164, 2013.

2. Jehad Ali, Rehanullah Khan, Nasir Ahmad, Imran Maqsood. Random Forests and Decision Trees. IJCSI International Journal of Computer Science Issues, vol. 9, 2012

3. Rajput, K., Oza B. A Comparative Study of Classification Techniques in Data Mining. International Journal of Creative Research Thoughts, vol. 5, pp. 154-163, 2017.

4. Shelke, P., Deshmukh, P. Person Identification Using Gait: SVM Classifier Approach. International Journal of Emerging Technologies and Engineering. vol. 1(10), 2014.

5. Johansson, U. Obtaining accurate and comprehensible data mining models: An evolutionary approach. Ph.D. dissertation, Institutionenf¨or datavetenskap, 2007.

6. Trifonov, R., Gotseva, D., Angelov, V. Analysis of data mining evaluation methods' efficiency. International Journal of Development Research, vol. 11(7), pp.16880-16884, 2017.