УДК 004.62 : [51-74 + 51-77]

27. ПРИМЕНЕНИЕ АЛГОРИТМОВ МАШИННОГО ОБУЧЕНИЯ И ОБОСНОВАНИЕ НЕОБХОДИМОСТИ ИХ ИСПОЛЬЗОВАНИЯ ПРИ АНАЛИЗЕ ГЕОГРАФИИ ПОТРЕБЛЕНИЯ БАНКОВСКИХ УСЛУГ КЛИЕНТАМИ

Юдашкин В. О., студент группы 172303

Белорусский государственный университет информатики и радиоэлектроники
г. Минск, Республика Беларусь

Мозоль А. А. – канд. экон. наук, доцент каф. ЭИ

Аннотация. В статье рассматриваются факторы, которые объясняют необходимость использования алгоритмов машинного обучения для анализа географии потребления услуг коммерческого банка его клиентами. Представлено описание практического применения алгоритма случайного леса, в том числе процессы обучения модели и прогнозирования примерных координат новых точек филиальной сети банка на основе тестовых данных, схожие с потенциальными реальными данными из базы коммерческого банка.

Ключевые слова. Алгоритм, искусственный интеллект, машинное обучение, анализ данных, прогнозирование, случайный лес.

При работе в банковской сфере в целом часто появляется необходимость большого объёма аналитики и взаимодействие с большими объёмами данных. Поэтому использование математических моделей и владение необходимой базой из математической статистики является неотъемлемым профессиональным навыком сотрудника банка. Применение методов и моделей регрессионного анализа, методов классификации и даже нейронных сетей помогает решать сложные задачи дескриптивной и предиктивной аналитики в банковской сфере.

Машинное обучение (Machine Learning – ML) – это раздел искусственного интеллекта, предоставляющий мощные математические модели для решения задач на основе больших объёмов данных и дальнейшего прогнозирования. В современном мире машинное обучение в рамках развития искусственного интеллекта постепенно внедряется во многие сферы деятельности человека. Однако сегодня банковская сфера является крупнейшим потребителем интеллектуальных технологий. Благодаря машинному обучению решаются следующие банковские задачи: принятие решений о выдаче кредита клиенту, анализируя его кредитную историю и поведение; выявление подозрительных операций и предотвращение мошенничества (fraud-аналитика); создание чат-ботов для взаимодействия с пользователями интернет-банкингов, используя модели обработки естественного языка (Natural Language Processing – NLP); прогнозирование курсов валют и рынка ценных бумаг, анализируя макроэкономические показатели.

В рамках разработки программного средства управления отделениями банка необходимо внедрение модуля анализа географии использования банковских услуг. Принято решение использовать алгоритмы машинного обучения для его реализации. Таким образом, были поставлены следующие цели: обоснование необходимости использования машинного обучения при анализе географии потребления услуг банка клиентами, определить факторные (входные данные) и целевые признаки (выходные данные), описать алгоритм машинного обучения, который будет применяться для выявления зависимости между входными и выходными данными, продемонстрировать качество модели при прогнозировании выходных данных.

География потребителей банковских услуг напрямую влияет на развитие филиальной сети банка, конкурентоспособность в разных регионах страны и качество предоставления услуг, так как спрос на отдельные услуги банка может резко отличаться, например, в крупных городах и агрогородках. Поэтому анализ географии использования банковских услуг является важной и неотъемлемой задачей в банковской сфере. Следует отметить, что крупные банки Республики Беларусь располагают тысячами клиентов в каждом административном районе страны, у которых между собой различаются уровень дохода и удовлетворённость услугами, что, в свою очередь, прямо влияет на открытие новых отделений банка. Возникает вопрос об анализе большой выборки клиентов, что приводит к значительным временным затратам. Таким образом, в целях оптимизации процесса анализа данных возникает необходимость использования алгоритмов машинного обучения, которые математически описывают даже самые неочевидные закономерности. В данном контексте естественный (человеческий) интеллект намного уступает искусственному.

В рамках разработки программного средства управления отделениями банка сформулирована задача машинного обучения, связанная с модулем анализа географии использования банковских услуг, цель которой – прогнозирование новых точек открытия отделений банка и (или) установки банкоматов. В качестве факторных признаков модели были взяты следующие показатели: *удовлетворённость клиентом услугами* (Customer Satisfaction – *S*) – данный фактор показывает оценку клиентом предоставления банком услуг, чем ниже эта оценка, тем больше вероятность, что он в будущем будет отдавать предпочтение банкам-конкурентам, что негативно сказывается на объёме потребления банковских услуг в регионе; *средний доход в регионе* (Average Income – *I*) коррелирует с объёмом потребления услуг в данном регионе, чем выше доход населения (потенциальных клиентов банка), тем активнее они потребляют банковские услуги; *плотность населения в регионе* (Population Density – *D*) имеет прямую зависимость непосредственно с показателями населения региона, чем больше плотность населения, тем больше может быть потенциальных клиентов банка, а соответственно, больший объём услуг банк может предоставлять в регионе; *количество отделений и банкоматов конкурентов в регионе* (Competitors – *C*) – чем больше конкурентов находится в регионе, тем меньший объём потребления услуг будет предоставлять банк, так как клиенты наиболее вероятно будут предпочитать конкурирующие банки.

В качестве целевых признаков модели взяты данные о местонахождении клиента – географические координаты, в частности его адреса прописки: *долгота* (Longitude – *Ln*) и *широта* (Latitude – *Lt*). Существует несколько подходов для решения поставленной задачи при помощи технологий машинного обучения. Это могут быть как простые модели регрессии, так и более сложные ансамблевые или нейросетевые модели. Одним из алгоритмов, который может решить поставленную задачу – это случайный лес (Random Forest), алгоритм которого схематически изображён на рисунке 1.

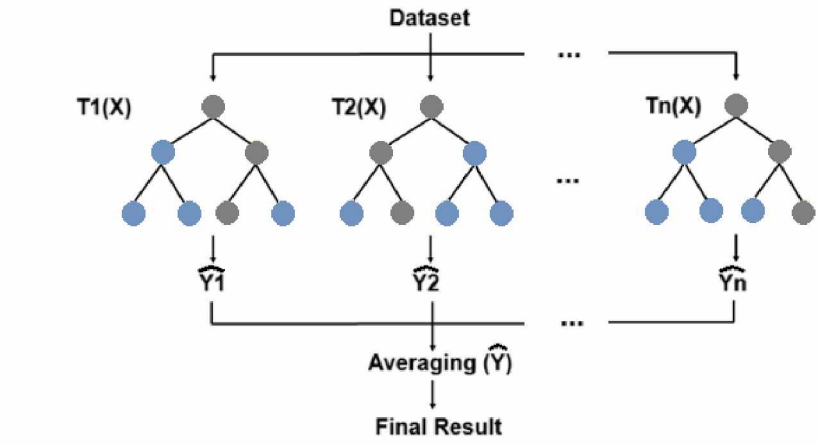

Рисунок 1 – Схема, демонстрирующая работу алгоритма «Случайный лес»

Случайный лес — это частный, оптимизированный случай бэггинга (параллельное обучение ансамбля из нескольких моделей) и вставки на основе решающих деревьев. Алгоритм строится следующим образом: изначально строится *n* выборок, как правило, методом случайных подпространств; *n* деревьев обучаются параллельно на полученных выборках и делают прогнозы на тестовой выборке; полученные прогнозы агрегируются модой в случае классификации и средним в случае регрессии.

Исследуемая задача машинного обучения имеет непрерывные данные в качестве выходных. Поэтому деревья решений строятся по алгоритму CART (Classification and Regression Tree) для поддержки решения регрессионных задач при помощи случайного леса. Модель обучается по следующей формуле:

$$\hat{Y} = \frac{1}{B}\sum_{b=1}^{B} T_b(X_b) \quad (1)$$

где $\hat{Y}$ – матрица прогнозных значений целевых признаков, X_b – матрица значений факторных признаков, охватывающая подпространство *b*, на основе которого обучалось дерево *b* для расчёта предсказаний каждого целевого признака, *B* – количество обучающихся решающих деревьев, $T_b(X_b)$ – результат работы дерева решений *b* над подпространством *b*.

Зная прогнозные значения целевых признаков, можно приступать к минимизации функции потерь *L*, основанной на среднеквадратической ошибке (Mean Squared Error – MSE):

$$\min_{W_{Lt},\, W_{Ln}} L(Y;\, \hat{Y}) = \frac{1}{nm}\sum_{i=1}^{n}\sum_{j=1}^{m}(y_{ij} - \widehat{y_{ij}})^2 \quad (2)$$

где *n* – это количество клиентов банка (объём выборки), *m* – количество целевых признаков (для данной задачи *m* = 2), иными словами, количество столбцов матриц *Y* и $\hat{Y}$.

За счёт минимизации функции потерь (среднеквадратической ошибки) были рассчитаны два вектора весов W_{Lt} и W_{Ln} для каждой географической координаты, которые передаются на вход функции регрессии вида *f(X; W)*. Теперь для расчёта прогнозного значения координаты необходимо воспользоваться формулой:

$$\widehat{Y_{n+1}} = \left(\widehat{Lt_{n+1}}; \widehat{Ln_{n+1}}\right) = \left(f(X_{n+1};\, W_{Lt}); f(X_{n+1};\, W_{Ln})\right) \quad (3)$$

где $(Lt_{n+1};\ Ln_{n+1})$ – предсказанное значение географической координаты (широта и долгота соответственно), X_{n+1} – вектор значений факторных признаков для строки *i*, W_{Lt} и W_{Ln} – вектора весов модели, полученные в результате минимизации функции потерь.

Для более наглядного представления применения описанной математической модели приведён контрольный пример. Представлена выборка из 10 клиентов (таблица 1), в дальнейшем будет выполнено прогнозирование местонахождения новых отделений банка на основе обучающей выборки. Входные и выходные данные в таблице являются тестовыми и используются только для проверки работы алгоритма случайного леса и демонстрации контрольного примера.

Таблица 1 – Исходные данные для демонстрации контрольного примера.

№ п/п	Количество отделений конкурентов	Средняя доходность в регионе (бел. руб./мес.)	Плотность населения в регионе (чел./км^2)	Удовл-ть услугами банка (от 0 до 100)	Местонахождение	
					Широта	Долгота
1	2	1407	11	59	53,31553	24,89556
2	0	1814	16	57	54,12300	25,84516
3	4	1646	25	68	54,33433	23,95625
4	5	2219	29	54	56,41589	27,56269
5	1	1896	18	49	52,51201	26,94800
6	1	1511	17	66	52,84150	26,85450
7	3	1356	20	37	51,99645	26,00052
8	1	1655	19	78	55,56222	27,85199
9	0	1700	10	68	53,46512	25,12121
10	2	1481	12	81	56,06520	24,96566

Модель случайного леса была написана на языке программирования Python. Она строится на основе входных и выходных данных, которые находятся в табличном формате *DataFrame*, который предоставляет библиотека pandas:

```
data = pd.DataFrame({
    "CustomerSatisfaction": customer_satisfaction,
    "PopulationDensity": population_density,
    "IncomeLevel": income_level,
    "Competitors": competitors,
    "Latitude": latitude,
    "Longitude": longitude
})
```

Ниже представлен листинг кода, который отвечает за непосредственное обучение модели. Можно обратить внимание, что выборка разделена на обучающую (80% выборки) и тестовую (20% выборки). Атрибут *n_estimators* хранит количество деревьев решений в лесу. Библиотека sklearn, предназначенная для решения задач машинного обучения в Python, представляет пакет для ансамблевых моделей sklearn.ensemble, который включает в себя алгоритм случайного леса (класс *RandomForestRegressor*):

```
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2, random_state=42)
model = RandomForestRegressor(n_estimators=100, random_state=42)
model.fit(X_train, y_train)
```

Далее проверяется прогнозирование на основе обученной модели. Выполняется предсказание координаты на основе следующих данных (таблица 2):

Таблица 2 – Данные для прогнозирования координаты.

Количество отделений конкурентов	Средняя доходность в регионе (бел. руб./мес.)	Плотность населения в регионе (чел./км^2)	Удовлетворённость услугами банка (от 0 до 100)
3	1899	18	70

На рисунке 2 представлен результат прогнозирования координаты:

```
Mean Squared Error: 0.81716
Predicted Latitude and Longitude: [[54.153617 25.963186]]
```

Рисунок 2 – Прогнозирование координаты для введённых данных.

Обученная ранее модель предсказала значения: 54.153617 северной широты, 25.963186 восточной долготы. Среднеквадратическая ошибка составила 0.81716 – значение, близкое к нулю, что говорит о том, что за счёт минимизации функции потерь были подобраны самые оптимальные значения весов, а значит модель работает качественно.

Предсказанная географическая координата не является конечной точкой, где необходимо открыть новый филиал банка, так как она может указывать на водоём, лес или другие природные препятствия. Так же географическая координата может указать на территорию соседнего государства, куда не распространяются услуги банка. Поэтому, используя представленный выше метод для решения поставленной задачи, необходимо ручное дополнительное исследование близлежащей местности и поиск на ней ближайшей доступной точки для открытия отделения банка или установки банкомата на территории населённого пункта.

Таким образом, был рассмотрен спектр задач в банковской сфере, в которой применяется машинное обучение, среди которых – задача анализа географии потребления услуг банка клиентами, которая была поставлена в рамках разработки программного средства управления отделениями банка, в которое должен быть внедрён данный аналитический модуль. Обоснована необходимость использования алгоритмов машинного обучения для решения поставленной задачи. Подробно продемонстрирован процесс решения задачи с помощью ансамблевого обучения на основе тестовых данных при помощи алгоритма «случайный лес», и, как выяснилось в результате предсказания данных,

задача решена успешно, а модель обучена качественно. Её среднеквадратическая ошибка варьируется в зависимости от исходных данных и их объёма, но для любого набора таких данных значение среднеквадратической ошибки остаётся близкой к нулю.

На основанной решённой задачи можно сформулировать следующие рекомендации:

1. Для тестирования и обучения модели с использованием тестовых данных генерировать их правдоподобными, чтобы модель училась идентично для тестовых и реальных банковских данных;

2. После выполнения прогнозирования географической координаты вручную выполнять перепроверку предсказанных данных, так как установка банкоматов и открытие филиалов возможно только на территории населённого пункта.

Следует акцентировать внимание на то, что решение задачи анализа географии использования банковских услуг с помощью алгоритма машинного обучения «случайный лес» является далеко не единственным. В случаях с несколькими выходными данными допускается обучение нейронной сети, однако для понимания её работы и применения необходимо углубленное изучение математических основ, согласно которым работает нейросеть. К тому же есть специальные алгоритмы машинного обучения, используемые в геоинформационных системах, прогнозирующие географические координаты и которые более эффективно могут решать данную задачу, в том числе сравнивая полученную географическую координату с местностью и выполнять перепроверку координаты при необходимости.

Список использованных источников:

1. Natural Language ToolKit [Электронный ресурс] – Режим доступа: https://www.nltk.org/

2. Бэггинг и случайный лес. Ключевые особенности и реализация с нуля на Python. [Электронный ресурс]. – Режим доступа: https://habr.com/ru/articles/801161/

3. scikit-learn Machine Learning in Python [Электронный ресурс] – Режим доступа: https://scikit-learn.org/stable/index.html