МРНТИ 4.1.5 https://doi.org/10.51488/1680-080X/2021.2-08
УДК 004.89, 004.912

С.З. Сапакова[1*], Н. Мадинеш [2]

[1] Международный Университет Информационных технологий, Алматы, Казахстан
[2] Казахский Национальный Университет им. аль-Фараби, Алматы, Казахстан
[*]Corresponding author: sapakovasz@gmail.com

Информация об авторах:
Сапакова Сая Заманбековна, к.ф.-м.н., ассистент-профессор, Международный Университет Информационных Технологий, кафедра Компьютерной инженерии и информационной безопасности, Алматы, Казахстан
https://orcid.org/0000-0001-6541-6806, email: sapakovasz@gmail.com
Мадинеш Н. – магистрант, Казахский Национальный Университет им. аль-Фараби, Алматы, Казахстан
https://orcid.org/0000-0001-9376-8489, aisymbat.9696@gmail.com

ПОЛУЧЕНИЕ ПРОГНОЗНЫХ ЗНАЧЕНИЙ ДЕМОГРАФИЧЕСКИХ ПРОЦЕССОВ С ИСПОЛЬЗОВАНИЕМ МОДЕЛЕЙ МАШИННОГО ОБУЧЕНИЯ

Аннотация. *Исследование и анализ демографических процессов играют важную роль во многих сферах. Для этого на статистическом сайте Республики Казахстан были выбраны численность населения и ключевые факторы с 1994 по 2019 год. Демографическими показателями были численность населения, рождаемость, смертность, разводы и миграция. Факторами уровня жизни были количество безработных и среднемесячная заработная плата, а медицинскими факторами – больничные организации, количество больничных коек и количество врачей всех специальностей. В ходе регрессионного анализа была получена корреляционная матрица и выявлены мультиколлинеарные факторы. Мы использовали четыре разные модели машинного обучения из библиотеки Scikit-Learn для получения оценок численности населения. Модели регрессии оценивались с использованием показателя качества R^2. В результате модели линейной регрессии и случайного леса показали хорошие результаты.*

Ключевые слова: *демографические процессы, регрессионный анализ, модели машинного обучения, показатели качества, библиотеки Scikit-Learn.*

Введение. В последние годы моделирование демографических процессов играет важную роль в исследованиях и во многих областях. Демографические данные Казахстана представляет собой обширный набор статистических данных о населении страны, включая такие разделы, как демография, здравоохранение, образование, уровень жизни, труд и занятости, социальное обеспечения и т.д. Анализируя данный вопрос, оптимальным будет использование современных технологиймашинного обучения.

Как мы знаем, машинное обучение занимается извлечением знаний из данных. Это научная область, находящаяся на пересечении статистики, искусственного интеллекта и компьютерных наук и также известная как прогнозная аналитика или статистическое обучение.

В статьи Cesare, Nina, Grant, Christan, Nguyen, Quynh, Lee, Hedwig, Nsoesie, Elaine O. [1] были исследованы и предложены методы определения демографических характеристик пользователей социальных сетей с использованием таких атрибутов, как имена пользователей и характеристика сети.

В статье Will Koehrsen [2] исследуются возможности Stocker, инструмента прогнозирования рынка, разработанного на Python. В Stocker прогноз производятся с использованием аддитивной модели, которая рассматривает временные ряды как комбинацию тренда и сезонных изменений в разных временных масштабах (ежедневный, еженедельный и ежемесячный).

В работе Jason Brownlee [3] исследуется набор данных о погоде и загрязнения в течение пяти лет в посольстве США в Пекине, Китай. В данной работе разрабатывается **модель LSTM для многомерного прогнозирования временных рядов** с помощью библиотеки глубокого обучения Keras.

В статье Aman Khakharia, Vruddhi Shah, Sankalp Jain, Jash Shah, Amanshu Tiwari, Prathamesh Daphal, Mahesh Warang, Ninad Mehendale [4] исследовали и разработали систему прогнозирования вспышек COVID-19 для лучших 10 густонаселенных стран.

Самая важная часть процесса машинного обучения – это интерпретация данных, с которыми мы работаем, и применимость этих данных к задаче, которую мы хотим решать.

Цель настоящей работы: Сравнение моделей машинного обучения при получении прогнозных значений демографических процессов.

Материалы и методы. Для достижения этой цели в ходе многофакторного регрессионного анализа изучаются следующие модели машинного обучения.

1. Регрессионный анализ – статистический метод, с помощью которого можно построить модель с одной зависимой переменной (откликом) и одной или несколькими независимыми переменными (факторами)[5].

2. Линейная регрессия (linear regression) является одним из самых простых обучающихся алгоритмов в нашем инструментарии [6].

3. Случайный лес (random forest) – это набор деревьев решений, где каждое дерево немного отличается от остальных. Идея случайного леса заключается в том, что каждое дерево может довольно хорошо прогнозировать, но, скорее всего, переобучается на части данных [7].

4. k-ближайших соседей (k-nearest neighbors) заключается предсказать метку нового объекта ¯x находятся k ближайших по расстоянию соседей этого объекта[8].

5. Метод опорных векторов (SVR) – это очень мощная и универсальная модель машинного обучения, способная выполнять линейную или нелинейную классификацию, регрессию и даже выявление выбросов [9].

Оценить качество регрессии можно таким же способом, которой мы использовали для классификации, например, сравнив количество завышенных и заниженных расчетных значений зависимой переменной. Однако в большинстве рассмотренных примеров будет достаточно применения R^2(«эр – квадрат»), который в методе score используется по умолчанию для всех моделей регрессии. Точки зрения оценки качества регрессионных моделей R^2(«эр – квадрат») является более понятной метрикой.

$$R^2 = 1 - \frac{RSS}{TSS}, \quad (1)$$

где

$$RSS = \sum_{t=1}^{n} e_t^2 = \sum_{t=1}^{n} (y_t - \hat{y})^2 \quad (2)$$

$$TSS = \sum_{t=1}^{n} (y_t - \bar{y})^2, \quad (3)$$

где y_t – истинное целевое значение; $\bar{y}$ – среднее значение вектора целей, для которого рассчитывается индикатор. Максимально доступные значения коэффициента обнаружения – 1.0, что соответствует идеальному состоянию прогнозной модели [1].

Результаты и обсуждения. Для анализа и разработки алгоритма прогнозирования демографических процессов использовалось программное обеспечение Jupiter. Для реализации алгоритма был использован язык программирования Python.

Анализ проводился на основе данных, полученных с этого статистического сайта Республики Казахстан. Как показано на Рисунке 2, население неуклонно растет. Факторы, влияющие на население, были выбраны на основе изучения демографических данных. На следующих рисунках выбираются наиболее важные факторы, и создается файл factornames.csv. Прогнозируемые значения совокупности были получены из этих данных с помощью регрессионного анализа и алгоритмов машинного обучения.

	date	population	births	deaths	divorces	Migration_gain	unemployed	Average_monthly_salary	hospitals	hospital_beds	doctors_of_all_specialties
0	1994	16942.00	305.62	160.34	41.57	400.90	70.08	1.73	1.65	205.7	61.1
1	1995	16679.00	276.13	168.66	38.65	330.00	80.00	4.79	1.52	192.6	60.1
2	1996	16544.00	253.18	166.03	40.50	270.00	282.41	6.84	1.24	164.4	57.9
3	1997	15993.00	232.36	160.14	35.74	242.64	257.48	8.54	1.01	136.4	54.5
4	1998	15804.00	222.38	154.31	35.46	220.00	251.94	9.68	0.99	123.5	53.2
5	1999	15000.00	217.58	147.42	25.58	210.52	251.38	11.26	0.92	108.2	50.6
6	2000	14901.64	222.05	149.78	27.39	324.14	231.40	11.80	0.94	106.9	49.0
7	2001	14865.61	221.49	147.88	29.60	325.28	216.10	57.00	0.98	110.2	51.3
8	2002	14851.06	227.17	149.38	31.24	327.30	193.70	60.00	1.01	111.9	53.7
9	2003	14866.84	247.95	155.28	31.72	357.34	142.80	63.00	1.03	114.8	54.6

Рисунок 1 – Населения Казахстана (1994-2019гг.)

При нахождении коэффициентов корреляции были выявлены мультиколлинеарные факторы. Развод(divorces) и количество врачей во всех сферах деятельности (doctors_of_all_specialties) были определены как факторы, оказывающие сильное влияние на население (population). Количество медицинских учреждений(hospitals) и больничных коек (hospital_beds) было определено как мультиколлинеарный фактор. Из факторов было определено p> 0,05, т.е. если значение p больше 0,05, мы удалили его из нашего анализа. При анализе коэффициент

детерминации составил R^2=0,99, т.е. качество модели было высоким. На следующем этапе мы разделили его на обучение и тестирование для изучения моделей машинного обучения. В результате были получены расчетные значения на рисунке 2.

	population	LinearRegression	RandomForestRegressor	KNeighborsRegressor	SVR
8	14851.06	15128.194963	15082.2081	15130.285000	15232.430643
24	18157.34	17616.846454	17650.7058	17126.600000	17369.530309
0	16942.00	16190.511236	16200.8807	16093.643333	16131.086148
22	17669.90	17433.045562	17485.2022	17126.600000	17264.964019
13	15396.88	15691.654026	15941.0907	15604.378333	15739.408257
17	16440.47	16493.247294	16235.3784	16278.018333	16369.947015
2	16544.00	16011.543320	15873.5378	15827.116667	15946.318131
23	17918.21	17580.152741	17650.7058	17126.600000	17329.188826
3	15993.00	15522.314259	15316.1199	15432.516667	15532.595258
6	14901.64	14677.259196	14983.9439	15093.736667	14824.334196
14	15571.51	15669.236371	15941.0907	15604.378333	15714.569863

Рисунок 2 – Фактические данные и прогнозные данные

Во время оценки моделей самые высокие значения были показаны линейная регрессия и случайный лес (рисунки 3, 4).

	TestModels	R2_score
0	LinearRegression	0.897533
1	RandomForestRegressor	0.890601
2	KNeighborsRegressor	0.830763
3	SVR	0.847407

Рисунок 3 – Оценка модели

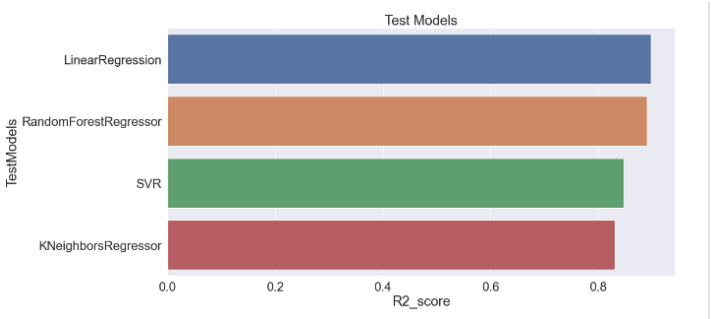

Рисунок 4 – Диаграмма оценки модели

Заключение. В нашем многофакторном регрессионном анализе мы определили развод и всех врачей как основные факторы, влияющие на население. В связи с этим модель получила высокую оценку. При исследовании моделей машинного обучения оценочные значения моделей линейная регрессия и случайный лес дали значения, близкие к заданному. Оценка моделей измерялась R^2(«эр − квадрат»). Мы проанализировали многофакторную регрессию с помощью библиотеки машинного обучения Scikit-Learn**.**

Литература:

1. *Cesare, Nina, Grant, Christan, Nguyen, Quynh, Lee, Hedwig, Nsoesie, Elaine O. How well can machine learning predict demographics of social media users?. 2017; 1702.01807v2.Available at: https://ui.adsabs.harvard.edu/abs/ 2017arXiv170201807C. Pp.1-24.*
2. *Will Koehrsen. Прогнозирование фондового рынка на Python с помощью Stocker//Neurohive;-2018.-URL:https://neurohive.io/ru/tutorial/ prognozirovanie-rynka-python-stocker/*
3. *Jason Brownlee. Multivariate Time Series Forecasting with LSTMs in Keras. – 2020. - URL: https://machinelearningmastery.com/multivariate-time-series-forecasting-lstms-keras/utm_source=dlvr.it&utm_medium=twitter*
4. *Aman Khakharia, Vruddhi Shah, Sankalp Jain, Jash Shah, Amanshu Tiwari, Prathamesh Daphal, Mahesh Warang, Ninad Mehendale. Outbreak Prediction of COVID-19 for Dense and Populated Countries Using Machine Learning //Annals of Data Science; 8(1):1–19. – 2021-URL:https://link.springer.com/content/pdf/ 10.1007/s40745-020-00314-9.pdf*
5. *Scott Robinson. Linear Regression in Python with Scikit-Learn. – 2020. – URL: https://stackabuse.com/linear-regression-in-python-with-scikit-learn/*
6. *Элбон Крис. Машинное обучение с использованием Python. Сборник рецептов: Пер. с англ. – СПб.: БХВ-Петербург, 2019. – 384 с.*
7. *Андреас Мюллер, Сара Гвидо. Введение в машинное обучение с помощью Python. Руководство для специалистов по работе с данными. Пер. с англ. – СПб.: БХВ-Москва, 2017. – 390 с.*
8. *Вьюгин В.В. Математические основы теории машинного обучения и прогнозирования. – М., 2013. – 387 с.*
9. *Жерон, Орельен. Прикладное машинное обучение с помощью Scikit-Learn и TensorFlow: концепции, инструменты и техники для создания интеллектуальных систем. Пер. с англ. – СПб.: ООО «Альфа-книга», 2018. – 688 с.*

References:

1. *Cesare, Nina, Grant, Christan, Nguyen, Quynh, Lee, Hedwig, Nsoesie, Elaine O. How well can machine learning predict demographics of social media users?. 2017; 1702.01807v2. URL: https://ui.adsabs.harvard.edu/abs/2017arXiv 170201 807C. Pp.1-24.*
2. *Will Koehrsen. Prognozirovaniye fondovogo rynka na Python s pomoshchyu Stocker//Neurohive;-2018.URL:https://neurohive.io/ru/tutorial/ prognozirovanie -rynka-python-stocker/*
3. *Jason Brownlee. Multivariate Time Series Forecasting with LSTMs in Keras.-2020.URL:https://machinelearningmastery.com/multivariate-time-series-forecasting-lstms-keras/utm_source=dlvr.it&utm_medium=twitter*
4. *Aman Khakharia, Vruddhi Shah, Sankalp Jain, Jash Shah, Amanshu Tiwari, Prathamesh Daphal, Mahesh Warang, Ninad Mehendale. Outbreak Prediction of COVID-19 for Dense and Populated Countries Using Machine Learning //Annals of Data Science; 8(1):1–19. – 2021.URL:https://link.springer.com /content/pdf/ 10.1007/s40745-020-00314-9.pdf.*
5. *Scott Robinson. Linear Regression in Python with Scikit-Learn. -2020.URL: https://stackabuse.com/linear-regression-in-python-with-scikit-learn/*
6. *Elbon Kris. Mashinnoye obucheniye s ispolzovaniyem Python. Sbornik retseptov: Per. s angl. – SPb.: BKhV-Peterburg. 2019. – 384 s.*
7. *Andreas Myuller. Sara Gvido. Vvedeniye v mashinnoye obucheniye s pomoshchyu Python. Rukovodstvo dlya spetsialistov po rabote s dannymi. Per. s angl. – SPb.: BKhV-Moskva, 2017. – 390 s.*
8. *Vyugin V.V. Matematicheskiye osnovy teorii mashinnogo obucheniya i prognozirovaniya. – M., 2013. – 387 s.*

9. *Zheron, Orelyen. Prikladnoye mashinnoye obucheniye s pomoshchyu Scikit-Learn i TensorFlow: kontseptsii. instrumenty i tekhniki dlya sozdaniya intellektualnykh sistem. Per. s angl. – SpB.: OOO «Alfa-kniga», 2018. – 688 s.*

С.З. Сапақова[1*], Н. Мәдінеш[2]

[1]Халықаралық Ақпараттық Технологиялар Университеті, Алматы, Қазақстан
[2] Әл-Фараби атындағы Қазақ Ұлттық Университеті, Алматы, Қазақстан
[*]Corresponding author: sapakovasz@gmail.com

Информация об авторах:
Сапакова Сая Заманбековна – физика-математика ғылымдарының кандидаты, Халықаралық Ақпараттық Технологиялар Университеті, Алматы, Қазақстан
https://orcid.org/0000-0001-6541-6806, email: sapakovasz@gmail.com
Мадинеш Н – Әл-Фараби атындағы Қазақ Ұлттық Университеті, Алматы, Қазақстан
https://orcid.org/0000-0001-9376-8489, aisymbat.9696@gmail.com

МАШИНАЛЫҚ ОҚЫТУ МОДЕЛЬДЕРІН ҚОЛДАНЫП ДЕМОГРАФИЯЛЫҚ ПРОЦЕССТЕРДІҢ БОЛЖАМДЫ МӘНДЕРІН АЛУ

Аңдатпа. *Көптеген салаларда демографиялық процестерді зерттеу мен талдау жасау маңызды рөл атқарады. Осы мақсатта Қазақстан республикасының статистикалық сайтынан 1994 жылдан 2019 жылға дейінгі халық саны мен негізгі факторлар іріктеліп алынды. Демографиялық көрсеткіштер ретінде халық саны, туу, өлім, ажырасу және көші-қон көрсеткіштері алынды. Өмір деңгейінің факторлары ретінде жұмыссыздар саны мен орташа айлық жалақы, ал медициналық факторлар ретінде аурухана ұйымдары, ауруханалардағы орын саны және барлық мамандық дәрігерлер саны алынды. Регрессиялық талдау жасау барысында корреляциялық қатынасы алынып, мультиколлинеарлық факторлар анықталды. Халық санының болжамды мәндерін алу үшін Scikit-Learn кітапханасыныі төрт түрлі машиналық оқыту модельдерін қолдандық. Регрессиялық модельдер R^2 сапа көрсеткішін қолдану арқылы бағаланды. Нәтижесінде сызықтық регрессия (linear regression) мен кездейсоқ орман(random forest) модельдері жақсы көрсеткіш берді.*

Түйін сөздер: *демографиялық процесстер, регрессиялық талдау, машиналық оқыту модельдері, сапа көрсеткіші, Scikit-Learn кітапханасы.*

Sapakova S.Z.[1*], Madinesh N.[2]

[1] International University of Information Technology, Almaty, Kazakhstan
[2] Kazakh National University named after Al-Farabi, Almaty, Kazakhstan
[*]Corresponding author: sapakovasz@gmail.com

Information about authors:
Sapakova Saya Zamanbekovna – candidate of physics –mathematic sciences, Assistant Professor, International University of Information Technology, Department of Computer Engineering and Information Security, Almaty, Kazakhstan
https://orcid.org/0000-0001-6541-6806, email: sapakovasz@gmail.com
Madinesh N. – Master's degree students al-Farabi Kazakh National University, Faculty of Information Technology, Almaty, Kazakhstan
https://orcid.org/0000-0001-9376-8489, aisymbat.9696@gmail.com

OBTAINING PREDICTED VALUES OF THE DEMOGRAPHIC PROCESS USING MACHINE LEARNING METHODS

Abstract. *Research and analysis of demographic processes play an important role in many areas. For this, the population size and key factors from 1994 to 2019 were selected on the statistical website of the Republic of Kazakhstan. Demographics were population size, fertility, mortality, divorce, and migration. The factors of the standard of living were the number of unemployed and the average monthly salary, while the medical factors were the hospital organizations, the number of hospital beds and the number of doctors of all specialties. In the course of regression analysis, a correlation was obtained and multicollinear factors were identified. We used four different machine learning models from the Scikit-Learn library to generate population estimates. Regression models were evaluated using the quality score R^2. As a result, linear regression and random forest models performed well.*

Keywords: *demographic processes, regression analysis, machine learning models, quality indicators, Scikit-Learn library.*