

Применение методов машинного обучения для построения модели прогнозирования цены холодильников в облачной среде Google Colaboratory и в системе Orange

Онохова Вера Алексеевна

Приамурский государственный университет имени Шолом-Алейхема

Студент

Аннотация

В данной статье описывается применение регрессионного анализа для прогнозирования цен холодильников. Для проведения исследования использовалась бесплатная среда Orange, написанная на языке Python, а также бесплатная облачная среда Google Colaboratory. Для примера собраны данные по холодильникам с сайта mvideo.ru. В результате исследования, путем создания моделей, удалось спрогнозировать стоимость покупки либо продажи холодильников точность до 89,2% по R-квадрату.

Ключевые слова: регрессионный анализ, метод ближайших соседей, линейная регрессия, Orange, python, Google Colaboratory.

Application of machine learning methods to build a model for predicting the price of refrigerators in the Google Colaboratory cloud environment and in the Orange system

Onokhova Vera Alekseevna

Sholom-Aleichem Priamursky State University

Student

Abstract

This article describes the use of regression analysis to predict refrigerator prices. To conduct the research, the free Orange environment written in Python was used, as well as the free Google Colaboratory cloud environment. For example, data on collected from the M.Video website. As a result of the research, by creating models, it was possible to predict the cost of buying or selling a laptop with an accuracy of up to 89,2% by R-squared.

Keywords: regression analysis, nearest neighbor method, linear regression, Orange, python, Google Coollaboratory.

1 Введение

1.1 Актуальность

В последние годы технологии машинного обучения находят применение в различных сферах. Одним из их направлений является предсказание цен. Это дает возможность оценить стоимость покупки или продажи конкретного товара. Для осуществления прогнозирования необходимо собрать

информацию по разным параметрам, из которых один будет состоять из цены, указанной в объявлении. На основе собранных данных можно создать различные модели, способствующие прогнозированию. Среди ключевых методов, используемых для предсказания, выделяется регрессионный анализ.

1.2 Обзор исследований

Е. Р. Клименко в своей статье «Разработка модуля прогнозирования цен на товары для онлайн-платформы» рассматривает современные подходы к прогнозированию цен, акцентируя внимание на важности использования алгоритмов машинного обучения для повышения точности прогнозов в электронной коммерции. Автор подчеркивает, что внедрение таких технологий может значительно улучшить процесс принятия решений для онлайн-торговцев [1].

И. С. Андриенко в работе «Применение методов машинного обучения для построения модели прогнозирования динамики цен на рынке автомобилей в среде Google Colaboratory» исследует, как доступные инструменты могут быть использованы для создания эффективных моделей прогнозирования. Он демонстрирует, что использование машинного обучения позволяет не только анализировать данные, но и строить предсказательные модели, что является актуальным для автомобильной отрасли [2].

А. А. Осин и его коллеги в статье «Использование методов машинного обучения для решения задач прогнозирования» обсуждают широкий спектр методов, применяемых для прогнозирования в различных областях. Они акцентируют внимание на том, что машинное обучение предоставляет мощные инструменты для анализа данных и решения сложных задач, что делает их работу полезной для исследователей и практиков [3].

В статье Т. В. Ермоленко, Д. В. Попадина и В. Н. Котенко «Применение машинного обучения в прогнозировании фондового рынка» рассматриваются особенности временных рядов финансовых данных и трудности их прогнозирования. Основная проблема заключается в стационарности и наличии памяти, что требует применения дробного дифференцирования для достижения стационарности. Авторы предлагают две модели машинного обучения: одну для маркировки движения цены по методу трёх барьеров и вторую для вычисления коэффициента уверенности с помощью двухступенчатого метода Лопеса Де Прадо [4].

Исследование Н. Бегларян и Д. В. Распопова «Прогнозирование цен с помощью методов анализа данных и машинного обучения» посвящено разработке модели для прогнозирования цен на электроэнергию в рамках проекта компаний «Росатома» РЭИН и «Цифрум». Авторы предполагают, что использование исторических данных о выработке электроэнергии и ценах на газ позволит прогнозировать цены на несколько суток вперёд, что повысит эффективность финансового планирования. Для этого они предлагают создать комплексную прогнозную модель на основе фактических цен и предикатов, использующую методы машинного обучения для краткосрочного прогнозирования [5].

1.3Цель исследования

Целью исследования является построение моделей прогнозирования цены подержанных холодильников на основе регрессионного анализа.

2 Материалы и методы

Для решения поставленной цели отлично подойдет бесплатная облачная среда Google Coollaboratory, разработанная для написания программного кода с использованием различных библиотек машинного обучения. А также бесплатная среда Orange, написанная на языке программирования Python.

3 Результаты и обсуждения

Сбор данных осуществлялся по данным сайта М.Видео 2024 года. Рассматривались холодильники разных годов выпуска по всей России. Данные собирались по нескольким параметрам, а именно: год выпуска, компания, высота, ширина, глубина, общий объем, энергоэффективность цена. Все было собрано в файл MS Excel (рис. 1).

Год	Модель	Компания	Высота (см)	Ширина (см)	Глубина (см)	Общий объем (л)	Энергоэффективность (кВтч)	Цена Log(Цена)
2023	Холодильник Gorenje NRK6202AW4 белый	Gorenje	200	60	59,2	353	248	54000
2023	Холодильник Candy CCRN 6200 S серебристый	Candy	200	59,5	65,7	409	395	40000
2020	Холодильник Indesit ITS 4180 W белый	Indesit	185	60	62	298	364	35000
2022	Холодильник Beko B3RCNK403ZW	Beko	201	59,5	65	403	276	40000
2024	Холодильник Candy CCRN6200WD	Candy	200	59,5	65,7	370	395	40000
2024	Холодильник HI HFDD019560SW белый	HI	195	60	65,8	344	343	30000
2020	Холодильник Candy CCRN 6180W белый	Candy	185	59,5	65,7	370	365	30000
2023	Холодильник Indesit ITS 5200 W белый	Indesit	196	60	64	325	379	41000
2023	Холодильник HI HCDN01885TDW белый	HI	188,4	57,4	62,5	329	302	28000
2020	Холодильник Candy CCRN 6200W белый	Candy	200	59,5	65,7	409	395	40000
2024	Холодильник Candy CCRN6200SD	Candy	200	59,5	65,7	409	395	47000
2024	Холодильник Haier C2F637CWMVU1	Haier	199,8	59,5	67,2	386	349	65000
2024	Холодильник Haier C2F636CWMVU1	Haier	190,5	59,5	67,2	364	342	65000
2023	Холодильник Candy CCRN 6180S серебристый	Candy	185	59,5	65,7	370	365	40000
2022	Холодильник Haier CEF531ASG	Haier	200	59,5	65	368	380	50000
2024	Холодильник Gorenje NRK620EABXL4	Gorenje	200	60	59,2	353	248	50000
2024	Холодильник с нижней морозильной камерой Grundig GKNP66930FXD	Grundig	201	59,5	65	357	276	50000
2023	Холодильник Haier CEF531ASG	Haier	190	59,5	65	346	360	54000
2024	Холодильник Haier C2F636CFDDU1	Haier	190,5	59,5	67,2	364	342	62000
2024	Холодильник Haier C2F637CWMVU1	Haier	199,8	59,5	67,2	386	349	67000
2024	Холодильник Haier CEF531AWG белый	Haier	200	59,5	65	368	380	57000
2024	Холодильник Haier C2F636CWFDDU1	Haier	190,5	59,5	67,2	364	342	65000
2023	Холодильник Haier CEF531AWG	Haier	190	59,5	65	346	360	54000
2024	Холодильник с нижней морозильной камерой Indesit ITS 5200 G	Indesit	196	60	62	353	377	39000
2023	Холодильник Haier C4F640CWMU1 белый	Haier	200	59,5	68,2	400	378	40000
2024	Холодильник ATLANT XM-4613-101	ATLANT XM	176,8	59,5	62,9	321	269	34000
2024	Холодильник Hotpoint HT 92011 MX O3	Hotpoint HT	200	60	62	357	323	58000
2021	Холодильник Gorenje NRK6202AXL4 серебристый	Gorenje	200	60	59,2	353	248	68000
2024	Холодильник Hotpoint HT 92011 W O3	Hotpoint HT	196	60	62	366	243	54000

Рисунок 1 – Фрагмент таблицы с данными MS Excel

Из таблицы видно, что все модели имеют довольно схожие характеристики. Для создания регрессионных моделей воспользуемся облачной средой Google Coollaboratory, а именно встроенными библиотеками pandas openpyxl scikit-learn matplotlib.

Проведя исследование для выявления наиболее подходящих параметров для анализа, выбрали такие показатели, как Год и Цена. В качестве основных методов анализа были выбраны линейная регрессия и метод ближайших соседей (kNN).

Для начала построим модель методом линейной регрессии. Так как один массив данных, а для проверки работы модели нужно два, то с помощью библиотеки pandas openpyxl scikit-learn matplotlib и модуля train_test_split создадим две выборки: обучающую и тестовую. И уже после этого с помощью модуля LinearRegression построим модель линейной регрессии по параметрам Год и Цена (рис.2).

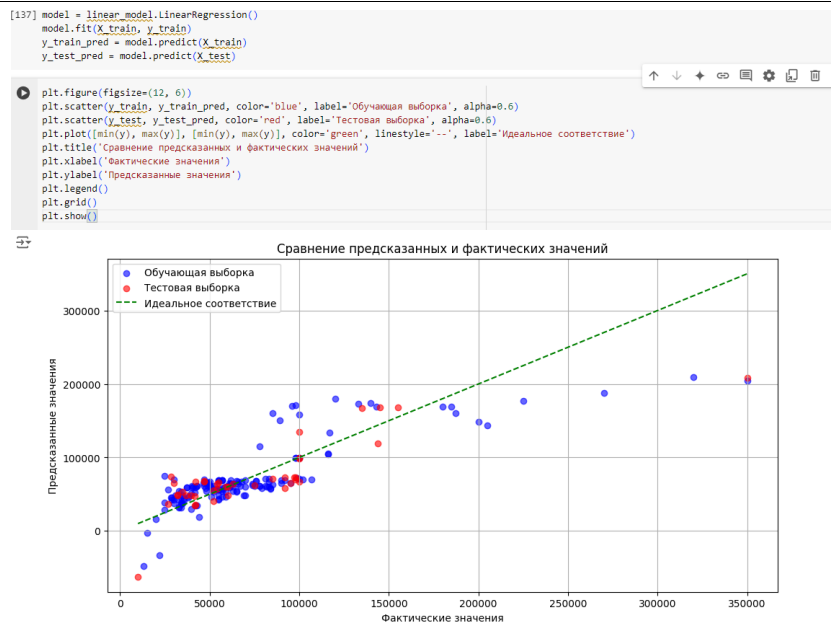


Рисунок 2 – Линейная регрессия (зависимость цены от года выпуска)

Полученный график представляет собой модель линейной регрессии. На рисунке видно, что точки обучающей выборки, то есть фактические и спрогнозированные цены, расположены значительно ближе к линии регрессии по сравнению с точками тестовой выборки. Это указывает на то, что модель лучше предсказывает значения для обучающей выборки, чем для тестовой.

Чтобы проверить хорошая ли модель получилась, снова воспользуемся модулями библиотеки pandas openpyxl scikit-learn matplotlib. Наиболее важным является `r2_score`, так как именно этот модуль поможет узнать коэффициент детерминации модели (R^2). Коэффициент детерминации для модели при правильном расчете принимает значения от 0 до 1. Для приемлемых моделей предполагается, что коэффициент детерминации должен быть хотя бы не меньше 50 %. У полученной модели линейной регрессии коэффициент детерминации составляет 69,1%, что говорит о том, что модель неплоха (рис.3).

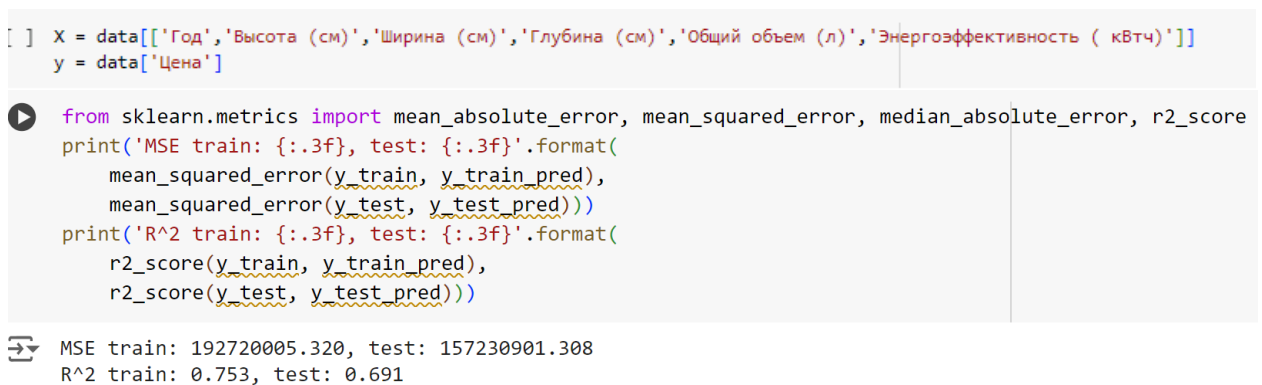


Рисунок 3 – Расчет коэффициента детерминации зависимости цены от года

После попробуем создать модель, используя все параметры: год выпуска, компания, высота, ширина, глубина, общий объем, энергоэффективность цена (рис.4).

```
▶ model_full = LinearRegression()  
model_full.fit(X, y)  
y_pred_full = model_full.predict(X)  
r2_full = r2_score(y, y_pred_full)  
print(f'Коэффициент детерминации общей модели: {r2_full}')
```

⇒ Коэффициент детерминации общей модели: 0.6930656640661992

Рисунок 4 – Расчет коэффициента детерминации общей модели линейной регрессии

На рисунке видно, что в данном случае модель вышла хуже, так как коэффициент детерминации равен уже 69,3%.

Но чтобы сделать распределение интенсивностей более симметричными, возьмем вместо параметра Цена десятичный логарифм от этого параметра и проведем повторное исследование с зависимостью всех параметров от логарифма от цены.

```
▶ y_pred_log = model_log.predict(X_test_log)  
r2_log = r2_score(y_test_log, y_pred_log)  
print(f'Коэффициент детерминации (логарифмированная цена): {r2_log}')
```

⇒ Коэффициент детерминации (логарифмированная цена): 0.7750303475397141

Рисунок 5 – коэффициент детерминации модели с логарифмированной ценой

И сразу заметен результата. Модель получилась достаточно достоверной, а коэффициент детерминации теперь равен 77,5%.

Также для прогнозирования цены используется модель созданная методом ближайших соседей (kNN). Для создания модели понадобятся параметры Год, и Цена.

С помощью библиотек pandas, scikit-learn, matplotlib. и модуля KNeighborsRegressor, а также используя все те же обучающую и тестовую выборки, создаем нужную модель, выводя как итог процент верно предсказанных вариантов.

```

▶ knn_3 = KNeighborsRegressor(n_neighbors=3)
  knn_3.fit(X_train, y_train)
  y_pred_knn_3 = knn_3.predict(X_test)
  r2_knn_3 = r2_score(y_test, y_pred_knn_3)
  print(f'Коэффициент детерминации (KNN, 3 соседа): {r2_knn_3}')

```

⇒ Коэффициент детерминации (KNN, 3 соседа): 0.32158620589712095

Рисунок 6 – Метод ближайших соседей (3 соседа)

На рисунке видно, что процент верно предсказанных вариантов составляет 32,1%, что является плохим показателем. Сделаем анализ со всеми параметрами, как и в прошлом методе (рис.6).

```

▶ knn_6 = KNeighborsRegressor(n_neighbors=6)
  knn_6.fit(X_train, y_train)
  y_pred_knn_6 = knn_6.predict(X_test)
  r2_knn_6 = r2_score(y_test, y_pred_knn_6)
  print(f'Коэффициент детерминации (KNN, 6 соседей): {r2_knn_6}')

```

⇒ Коэффициент детерминации (KNN, 6 соседей): 0.5060327892401513

Рисунок 6 – Метод ближайших соседей (6 соседей)

В данном случае модель стала немного лучше. Добавим снова десятичный логарифм от цены, чтобы сравнить результат (рис.7).

```
[107] knr6 = KNeighborsRegressor(n_neighbors=6)
```

```

▶ knr6.fit(x2_train, y2_train)

  train_score = knr6.score(x2_train, y2_train)
  print(f'Training score: {train_score:.3f}')

```

⇒ Training score: 0.777

Рисунок 7 - Метод ближайших соседей (6 соседей)

Модель стала лучше, процент успешно предсказанной цены составляет 77,7%, что говорит о том, что модель достаточно достоверна.

Похожий анализ также был проведен с помощью среды Orange, но с использованием большим количеством регрессионных моделей.

Внеся собранные данные в Orange и разделив данные на тестовые и тренировочные в пропорции 80/20 с помощью объекта Date Sampler и

использую кроссвалидацию с разбиением на 5 фолдов, а также добавив ещё несколько объектов: Distributions , Correlations, Scatter Plot. Свяжем их с Test and Score (рис. 8). Уже от Test and Score проведем к нужным моделям, выставив верные параметры. В качестве моделей для прогнозирования выбраны Random Forest, Gradient Boosting, kNN, AdaBoost и Linear Regression.

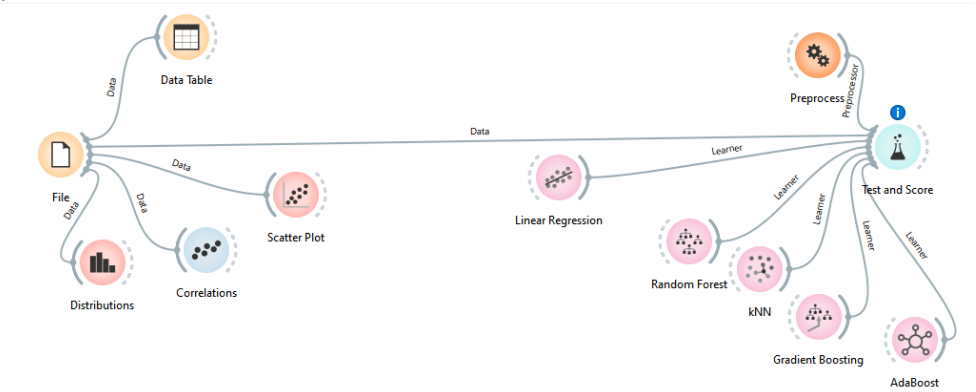


Рисунок 8 - Полная модель

Программная среда Orange самостоятельно проводит анализ, на основе внесенных данных и выставленных параметров. Результат проведенного анализа выводится в Test and Score в виде коэффициента детерминации, из которого уже можно сделать вывод о качестве каждой модели (рис.9).

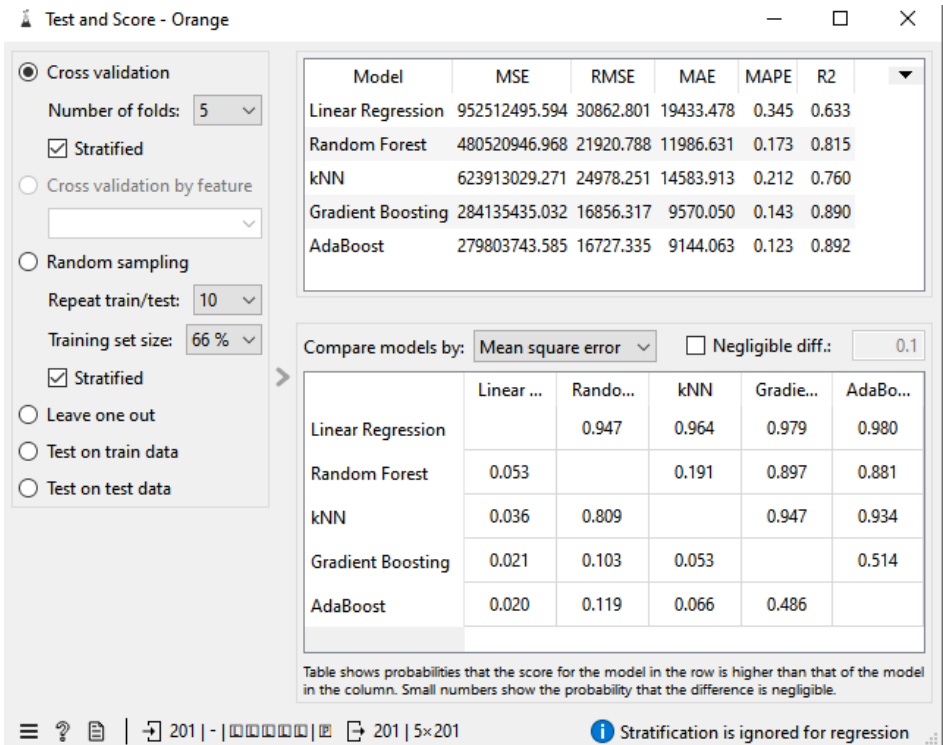


Рисунок 9 - Результат в Test and Score

Самой достоверной и хорошей из моделей получилась модель AdaBoost, коэффициент детерминации модели с логарифмированной ценой 89,2%.

Google Colaboratory. - URL:

https://colab.research.google.com/drive/18uBNkrUA_L032tNjTA0pqVSJzisR1UPV?usp=sharin

Данные. - URL: <https://disk.yandex.ru/d/BxnfwSFCh-IkRQ>

Таким образом, полученные модели прогнозирования, созданные в Google Colaboratory и Orange можно использовать для покупки холодильников разных годов выпуска, а данные при прогнозе будут достаточно достоверными. Проведенное исследование может быть использовано при обучении методам машинного обучения в различных направлениях.

Библиографический список

1. Клименко Е. Р. Разработка модуля прогнозирования цен на товары для онлайн-платформы // Молодой ученый. 2024. № 41 (540). С. 5-6. URL: <https://moluch.ru/archive/540/118129/> (дата обращения: 03.01.2025).
2. Андриенко И. С. Применение методов машинного обучения для построения модели прогнозирования динамики цен на рынке автомобилей в среде Google Colaboratory // Постулат. 2024. №6. URL: <https://e-postulat.ru/index.php/Postulat/article/view/5688> (дата обращения: 13.01.2025).
3. Осин А. А., Фомин А. К., Сологуб Г. Б., Виноградов В. И. Использование методов машинного обучения для решения задач прогнозирования // Моделирование и анализ данных. 2020. Т. 10. №4. URL: <https://cat.gpntb.ru/?id=EC/ShowFull&irbDb=KATBW&bid=-995169-524368> (дата обращения: 13.01.2025).
4. Ермоленко Т. В., Попадин Д. В., Котенко В. Н. Применение машинного обучения в прогнозировании фондового рынка // Problems of Artificial Intelligence, 2023. № 2 (29). URL: http://paijournal.guide.ru/download_pai/2023_2/2_Ермоленко_Попадин_Котенко.pdf (дата обращения: 13.01.2025).
5. Бегларян Н., Распопов Д. Предсказываем цены с помощью методов анализа данных и машинного обучения // Блог компании «Росатом» 2023. URL: <https://habr.com/ru/companies/rosatom/articles/771534/> (дата обращения: 13.01.2025).
6. Шамраева В. В. Математические методы прогнозирования изменения цены акций и их реализация методами машинного обучения // Фундаментальные исследования. 2024. URL: <https://s.fundamental-research.ru/pdf/2024/11/43718.pdf> (дата обращения: 06.01.2025).
7. Андриенко И. С. Применение методов машинного обучения для построения модели прогнозирования динамики цен на рынке автомобилей в среде Google Colaboratory // Постулат. 2024. № 6 (14). С. 5-9. URL: <https://e-postulat.ru/index.php/Postulat/article/view/5688> (дата обращения: 06.01.2025).
8. Волоцкой Д. В. Применение методов машинного обучения для построения

- прогнозов на финансовых рынках. Национальный исследовательский университет ИТМО, г. Санкт-Петербург. 2022. URL: <https://админ.авсэ.пф/Files/ArticleFiles/9848a6e3-6c9f-47cb-b782-486c27d29383.pdf> (дата обращения: 06.01.2025).
9. Алжеев А. В., Кочкаров Р. А. Сравнительный анализ прогнозных моделей ARIMA и LSTM на примере акций российских компаний // Финансы: теория и практика. 2020. URL: <https://ideas.repec.org/a/scn/financ/y2020i1p14-23.html> (дата обращения: 03.01.2025).
10. Google Colaboratory. -
URL: https://colab.research.google.com/drive/18uBNkrUA_L032tNjTA0pqVSJzisR1UPV?usp=sharing