**Анализ результатов моделирования**

В процессе моделирования было построено две основные модели и несколько промежуточных в ходе их улучшения и подбора гиперпараметров. Промежуточные результаты моделей и оценки по ним представлены в таблице.

Для улучшения модели были произведены следующие действия.

1. Избавились от пропущенных значений в данных, заменив на минимум по датафрейму.

2. Избавились от переменных, оказывающих отрицательное влияние на качество модели.

3. Подобрали оптимальные гиперпараметры.

Далее рассмотрим подробнее две итоговые модели.

Первая модель предполагает использование только градиентного бустинга CatBoost. Рассматриваемый алгоритм удобен тем, что он не требует нормализации данных и слабо чувствителен к выбросам, поэтому переменные не подвергались предварительной обработке.

Вторая модель представляет из себя стекинг-ансамбль из CatBoost регрессии и линейной ридж-регрессии.

Из пробной модели с параметрами глубины дерева =3 и темпа обучения=0.1 с помощью метода выбора признаков из библиотеки CatBoost, основанного на алгоритме рекурсивного исключения признаков по SHAP-Values.



Рисунок 16 – Результат отбора признаков и их оценки

Таким образом, из датасета было исключено 12 признаков, а именно cpi, youthpop, condig, internetorg, citydig, roads, investment, citypop, sctec, incomepop, cargo.

Оптимальная модель по всем оценкам была построена по следующим параметрам с помощью встроенного метода поиска по сетке из библиотеки CatBoost: глубина дерева =7 и темпа обучения=0.5 (модель №2). На рисунке 17 представлен график результатов оценок RMSE подобранных моделей.

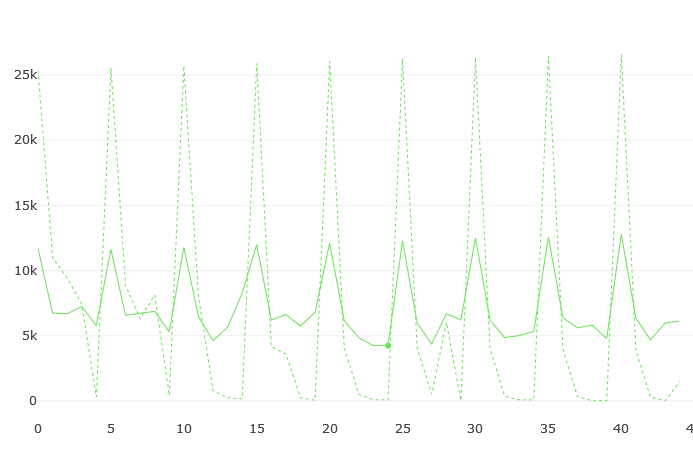


Рисунок 17 – График поиска по сетке оптимальных гиперпарметров модели

По этим же параметрам была построена модель-ансамбль, так как для обучения она принимает в себя ту же самую обучающую выборку. Также для ридж-регрессии был подобран параметр альфа =1.

В рассматриваемых моделях №2 и №4 были выбраны следующие признаки: card, covid, grp, poorpop, entrepreneur, internetpop, students.

Оценки модели представлены в таблице 1.

Таблица 1 – Сравнение оценок итоговых моделей

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | CatBoost | CatBoost+Linear | CatBoost | CatBoost+Linear |
| № | Model2 | Model4 | Model2 | Model4 |
| Выборка | train | | test | |
| Параметры модели | iteration: 54  depth: 7  learning\_rate: 0.5 | | iteration: 3179  depth: 7  learning\_rate: 0.5  alpha (ridge): 1 | |
| Оценки | | | | |
| MAE | 585.8895 | 5497.9049 | 8703.7902 | 6837.5675 |
| MAPE | 7.6486 | 1.4232 | 0.7516 | 0.9622 |
| SMAPE | 55.7586 | 125.6982 | 63.0250 | 60.9385 |
| RMSE | 650 536.7041 | 162 664 697.9054 | 359 861 892.8238 | 147 716 906.8992 |
| R2 | 0.9993 | 0.8421 | 0.9288 | 0.9708 |
| CV-score (RMSE) | 136.688 | 170.618 |  |  |
| Сумма esales за 2021 г, млн руб., факт | 2 005 482.1600 | | | |
| Сумма esales за 2021 г, млн руб., прогноз | - | - | 1 402 796.9598 | 1 964 750.6646 |
| Ошибка прогноза, млн руб. | - | - | -602 685.2002 | -40 731.4954 |

Обе модели между собой в основном отличаются по оценкам по обучающей и тестовой выборках.

Модель №2 лучше показывает себя на обучающей выборке, в то время как на тестовой выборке результат гораздо хуже.

В модели №4 все наоборот. На новых данных, которые не участвовали в обучении, модель показывает оценки лучше.

Также стоит отметить, что разница между оценками на обучающей выборке и на тестовой в модели №2 гораздо выше, чем у модели №4. Например, оценка MAE по тестовой выборке на 1385,569% больше, чем на обучающей. Данный факт, скорее всего, свидетельствует о том, что модель №2 имеет гораздо меньшее смещение, но при этом возрастает чувствительность к шуму и колебаниям новых данных.

Модель №4 позволяет сгладить некоторые сильные колебания и добиться большей точности прогноза: R2 удалось улучшить до 97%, а ошибку прогноза продаж уменьшить до 40 731,45 млн. руб.

Далее рассмотрим влияние признаков на целевую переменную через расчет значений Шэпли.

На рисунке 18 отображены усредненные значения Шэпли по каждому признаку по модулю. Наибольшее влияние оказывает показатель card, наименьшее – student.

Также рассмотрим рисунок 19, на котором отображен график значений Шэпли по всем наблюдениям. По всем признакам можно наблюдать четкие зависимости. По признакам card, grp, entrepreneur, internetpop, students можно сделать следующие выводы: чем выше значения целевой переменной esales, тем выше влияние признака. По переменной covid можно наблюдать, что сильным влиянием на целевую переменную оказывают значения по covid=1. По показателю poorpop можно наблюдать обратную зависимость: чем больше процент poorpop, тем слабее он влияет на целевую переменную.

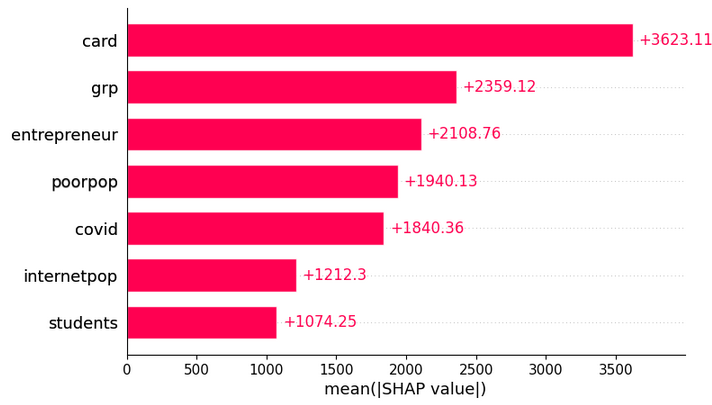


Рисунок 18 - усредненные значения Шэпли по каждому признаку по модулю по модели №2

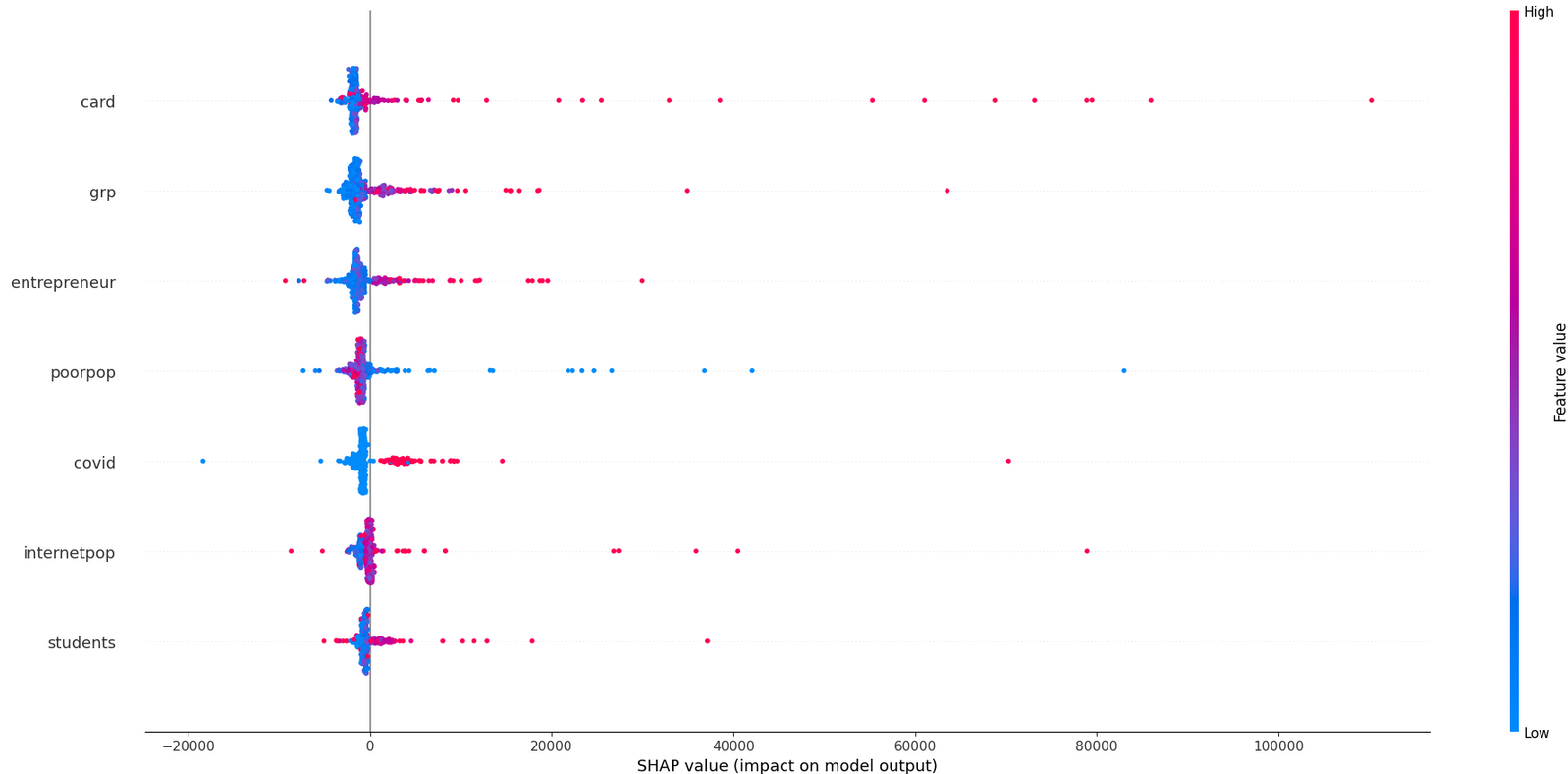


Рисунок 19 - график значений Шэпли по всем наблюдениям

На рисунке 20 отобразим распределение между значением переменной и ее значением Шэпли.

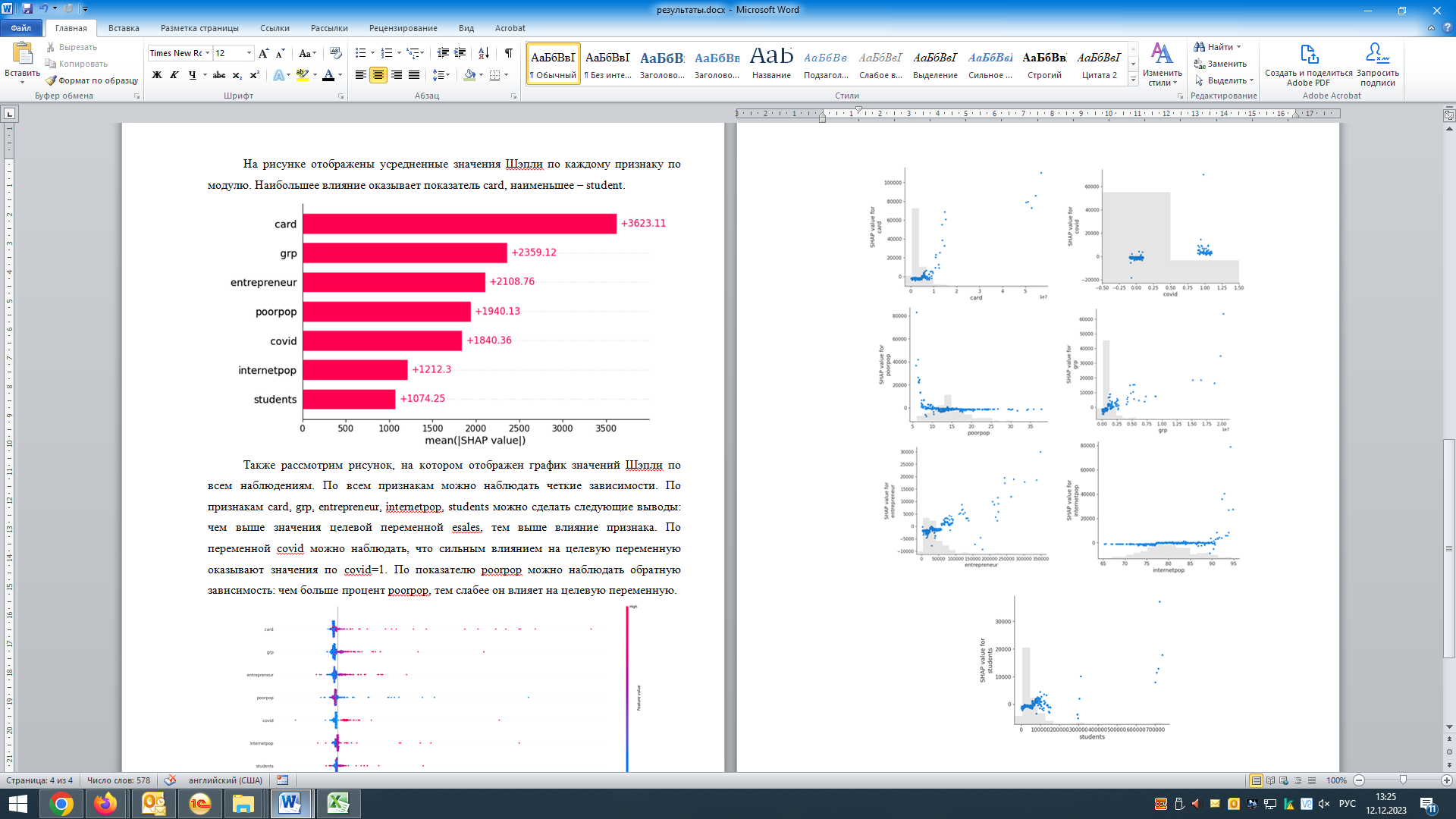


Рисунок 20 - распределение между значением переменной и ее значением Шэпли по всем признакам

По всем переменным прослеживаются слабые прямые зависимости (кроме poorpop: имеет обратную зависимость). Как уже ранее отмечалось, чем выше значения признака, тем сильнее его воздействие на целевую переменную.

Таким образом, с помощью градиентного бустинга CatBoost и его комбинации с линейной регрессией удалость построить прогностическую модель, которая на 97% описывает дисперсию целевой переменной интернет-продаж.

**Обсуждение результатов**

Определим признаки, которые оказывают влияние на интернет-продажи в РФ.

Наиболее значимым оказался показатель количества выданных карт. Он в большей степени прямопропорционально влияет на развитие рынка интернет-торговли, так как наличие карты позволяет дистанционно оплачивать покупки. Поэтому этот фактор можно считать основным драйвером рассматриваемой отрасли. Данный результат оказался ожидаемым.

Вторым по значимости является показатель внутреннего регионального продукта (ВРП). Чем экономически развитие регион в целом, тем выше у него будут интернет-продажи. Данный результат оказался ожидаемым.

Третьим по степени значимости оказался показатель количества предпринимателей в регионе. Чем больше количество предпринимателей в регионе, тем больше вероятность того, что они будут вести торговлю через Интернет, так как это гораздо проще, чем традиционный способ.

Следующим по значимости оказался показатель доли бедного населения. Он обратнопропорционально влияет на интернет-продажи. Для совершения покупок через Интернет необходим смартфон/компьютер с подключением к Интернету. Бедные слои населения не могут себе этого позволить, поэтому чем выше доля бедного населения, тем ниже интернет-продажи и, скорее всего, в общем розничные продажи.

Пандемия коронавируса также оказалась в числе воздействующих на интернет-продажи факторов. Но данный показатель скорее является не фактором, а одномоментным событием, которое лишь на время увеличило темпы развития рынка. Из-за локдауна в 2020-2021 гг. люди вынуждены были находиться дома и вести деятельность удаленно, в том числе совершать покупки. На данный момент население уже адаптировалось к этому. И скорее всего далее данный фактор перестанет быть значимым.

Также показатель распространенности Интернета среди населения со временем будет терять свою значимость за счет того, что данная технология будет считаться обыденностью, и каждый человек будет иметь выход в Интернет и доступ к интернет-магазинам.

Последний по значимости среди отобранных факторов является количество студентов в учреждениях высшего образования. Образование населения прямопропорционально влияет на совершение покупок через Интернет, так как для этого необходимо определенные навыки и знания. Люди с высшим образованием чаще сейчас сталкиваются с Интернетом и покупками через него, так как, скорее всего, их работа с большей вероятностью связана с ним.

Далее проанализируем результаты построенных моделей и используемых методов.

Почти во всех источниках из литературного обзора утверждается, что методы машинного обучения показывают лучшие результаты в сравнении с традиционными методам в области прогнозирования интернет-продаж. С этими выводами можно согласиться, так как полученные результаты данного исследования свидетельствуют о высоком качестве моделей машинного обучения, а именно градиентного бустинга. В ходе работы коэффициент детерминации R2 всегда оставался на уровне 83-90%.

Помимо этого, некоторые исследователи в своих работах демонстрируют высокие результаты комбинированных ансамблевых моделей. С их утверждениями можно частично согласится, так как благодаря применению линейной регрессии вместе с градиентным бустингом получилось снизить чувствительность модели к колебаниям новых входных данных. Поэтому удалось достигнуть уровня 97% коэффициента детерминации на тестовой выборке и избавиться от переобучения модели, но при этом снизив ее сложность и потеряв точность на обучающей выборке.

Таким образом, выдвинутая гипотеза исследования считается подтвержденной, так как отобранные показатели оказались статистически значимыми. Также был продемонстрированы преимущества моделей машинного обучения, а именно градиентного бустинга CatBoost и метода ансамбля, так как они показали хорошие высокие результаты оценок как на обучающей, так и на тестовой выборке.