

ПРИЛОЖЕНИЯ

Приложение 1

1 этап. Получение исходной выборки.

Информация о закупках, в которых участвовал картель, была получена из решений по делам ФАС об антиконкурентных соглашениях за период с 2014 по 2016 гг. Для извлечения номеров закупок использовались методы машинного обучения. Исходная выборка включила 214 дел, по которым признавалось наличие антиконкурентных соглашений на 5055 закупочных процедурах. Для контрольной группы закупок было случайным образом отобрано 20 000 аукционов из всего массива закупок в 2014–2016 гг.

2 этап. Предварительная обработка.

Далее была проведена предварительная обработка данных, в ходе которой были отброшены закупки с номерами, не соответствующими формату электронного аукциона, и дублирующие номера закупок. В результате осталось 18 015 (90,1%) аукционов случайной выборки и 4982 (98,6%) выборки по делам ФАС.

3 этап. Построение матрицы документов и фильтрация по шаблону.

Каждый этап электронного аукциона должен сопровождаться публикацией на сайте электронной площадки соответствующих документов: извещение, три протокола и итоги аукциона с указанием выбранного поставщика. Но в реальности некоторые документы часто отсутствуют или являются пустыми. В некоторых случаях есть, например, только извещение. Это может говорить о том, что аукцион был отменен из-за того, что не набралось минимальное необходимое число участников, либо заказчик передумал совершать закупку, либо были выявлены какие-то нарушения со стороны заказчика и т.д. В исходную выборку включились аукционы, для которых был опубликован хотя бы один из необходимых документов.

Нам интересны аукционы, которые успешно завершились, и в результате которых был определен поставщик, поэтому в итоговый набор данных были включены закупки содержащие:

- 1) извещение о проведении электронного аукциона (fcsNotificationEF);
- 2) протокол рассмотрения заявок на участие в электронном аукционе (fcsProtocolEF1);
- 3) протокол проведения электронного аукцион (fcsProtocolEF2);
- 4) протокол подведения итогов электронного аукциона (fcsProtocolEF3);
- 5) результат процедуры определения поставщика (fcsPlacementResult).

В результате фильтрации данных осталось 10 381 (51,9%) закупок из случайной выборки и 3452 (68,3%) – из выборки ФАС. Если аукцион успешно завершился, то по нему должны быть опубликованы минимум пять основных документов, указанных выше. Но по разным причинам число их может оказаться больше. Чаще всего публикуются дополнительные документы с разъяснениями положений о закупке. Реже публикуются извещения или протоколы.

4 этап. Извлечение данных из объединенного XML-документа.

На последнем этапе из документов, опубликованных на сайте ЕИС, извлекались отдельные показатели. Из Извещения была взята информация о начальной максимальной цене аукциона. Протокол рассмотрения заявок (1 Протокол) послужил источником данных о числе фирм-участников и о времени подачи заявок. Также была получена информация о том, какие из заявок были отклонены по первой части. Наиболее важным документом для нас был Протокол проведения аукциона

(2 Протокол), из которого были получены данные о величине ставок, их числе, длительности проведения аукциона и об итоговом снижении цены. Из Протокола подведения итогов (3 Протокол) была взята информация о том, какие фирмы были отклонены по второй части заявок, и кто был выбран победителем торгов (Не всегда совпадали тот, кто предложил наименьшую цену, с тем, кого в итоге выбрали поставщиком).

В некоторых закупках Протоколы оказались пустыми, соответствующие наблюдения были исключены из выборки. Также были исключены закупки с явными ошибками в документах, в частности: закупки, для которых снижение цены в долях от НМЦК меньше 0, близко к 1 или больше 1; разница между минимальной и второй ценой в долях от НМЦК больше 1; разница между временем подачи заявки и публикацией извещения меньше 0.

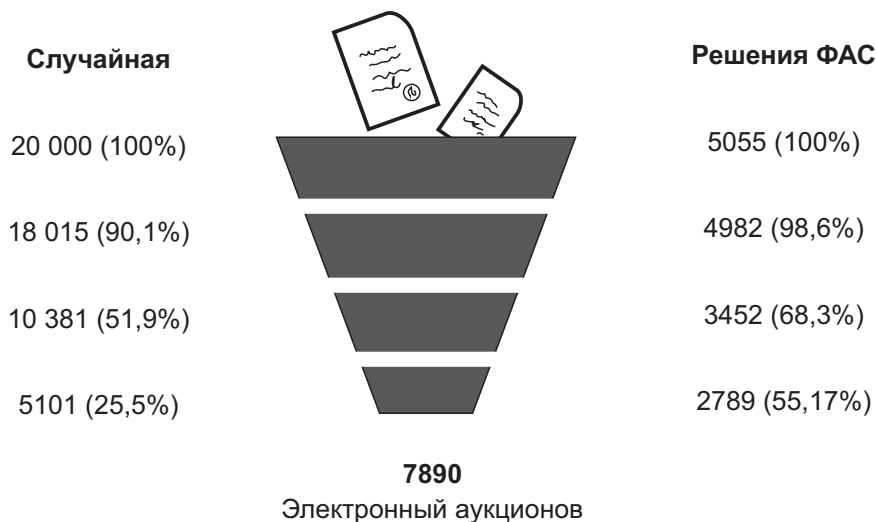


Рис. 10. Схема сужения выборки в результате фильтрации данных

Источник: расчеты авторов.

Приложение 2

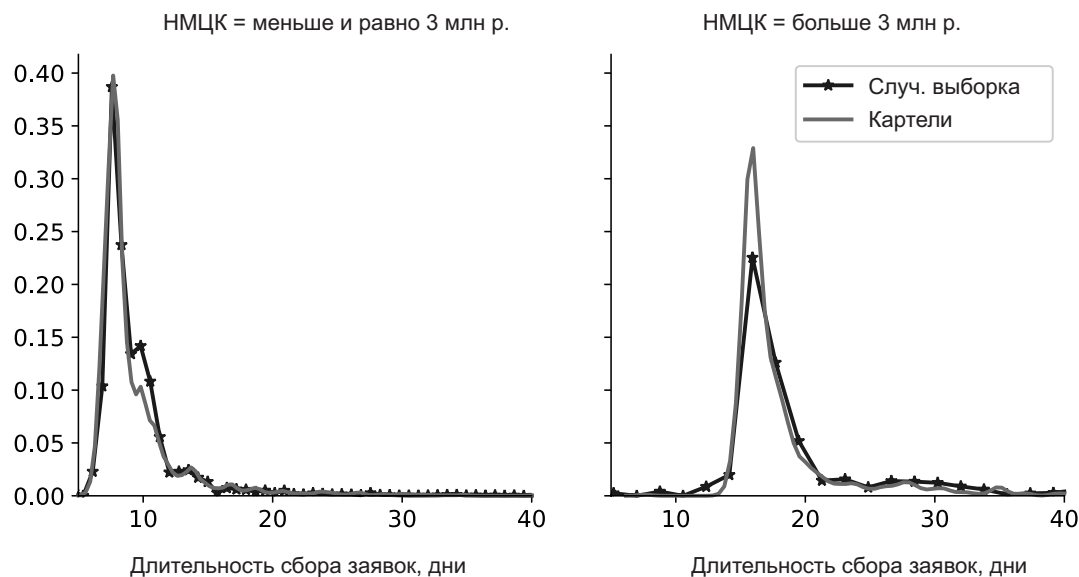


Рис. 11. Распределение длительности сбора заявок

Источник: рассчитано авторами.

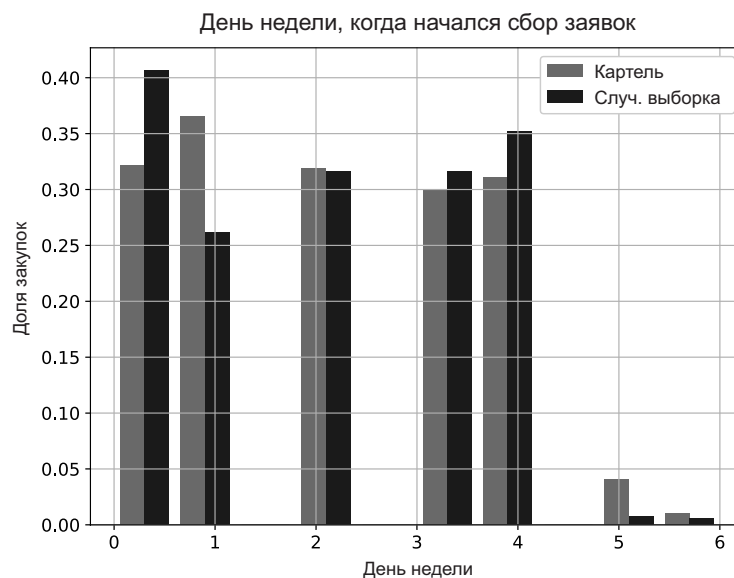


Рис. 12. Распределение времени начала сбора заявок по дням недели

Примечание: 0 – понедельник, 1 – вторник, 2 – среда, ... 6 – воскресенье.

Источник: рассчитано авторами.

Приложение 3

Для того чтобы наглядно показать совместные распределения разных переменных были построены диаграммы рассеяния для попарно взятых переменных (графики вне главной диагонали) и графики плотности распределения отдельных переменных (на главной диагонали) (рис. 13). Поскольку переменных много, то целесообразно разделить матрицу из всех парных графиков на несколько частей, чтобы была возможность изучить каждый элемент. Рисунок 13 был разделен на четыре части. Получившиеся 1, 2 и 3 четверти представлены на рис. 14–16, соответственно.

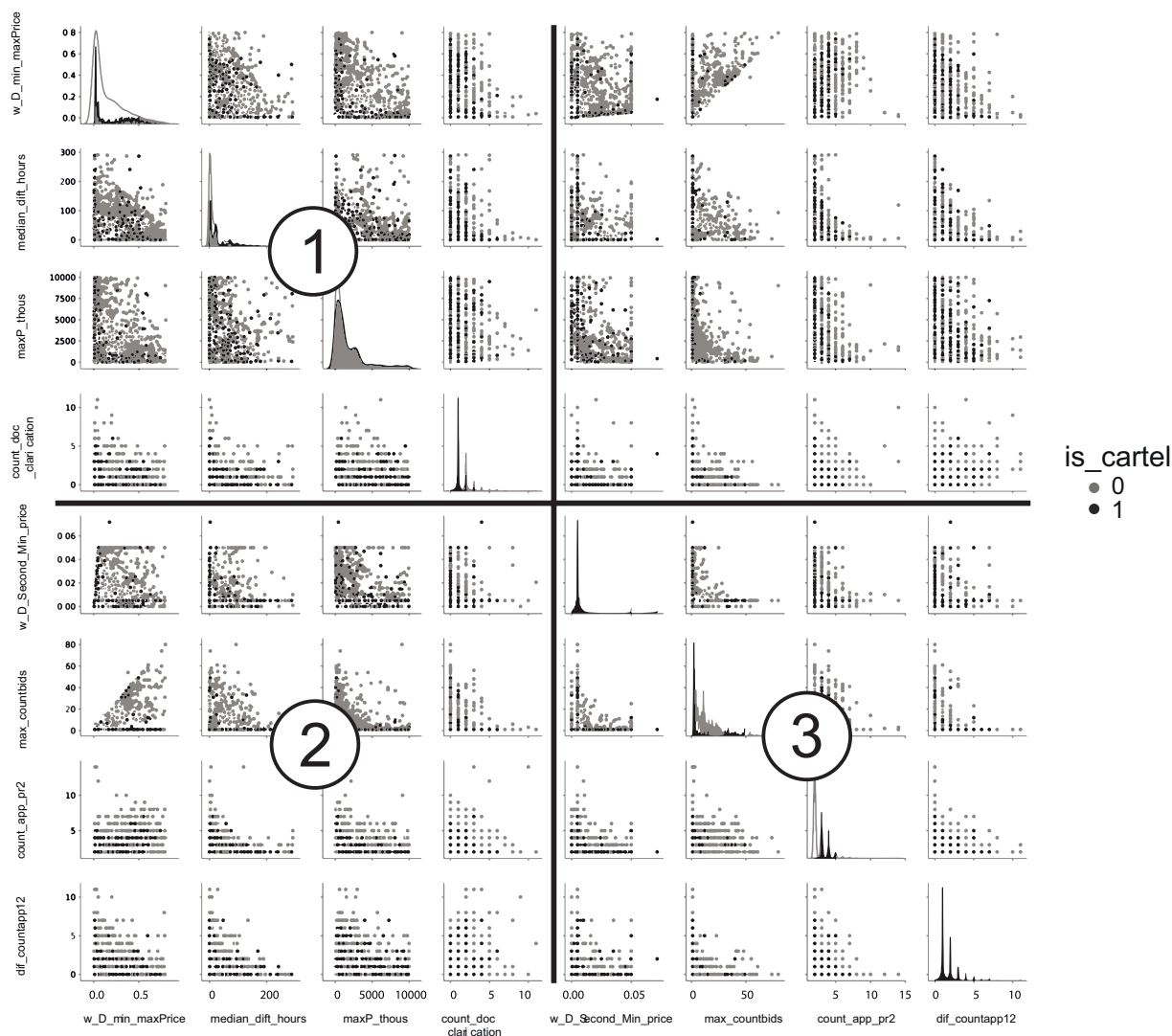


Рис. 13. Зависимости между всеми парами переменных

Источник: рассчитано авторами.

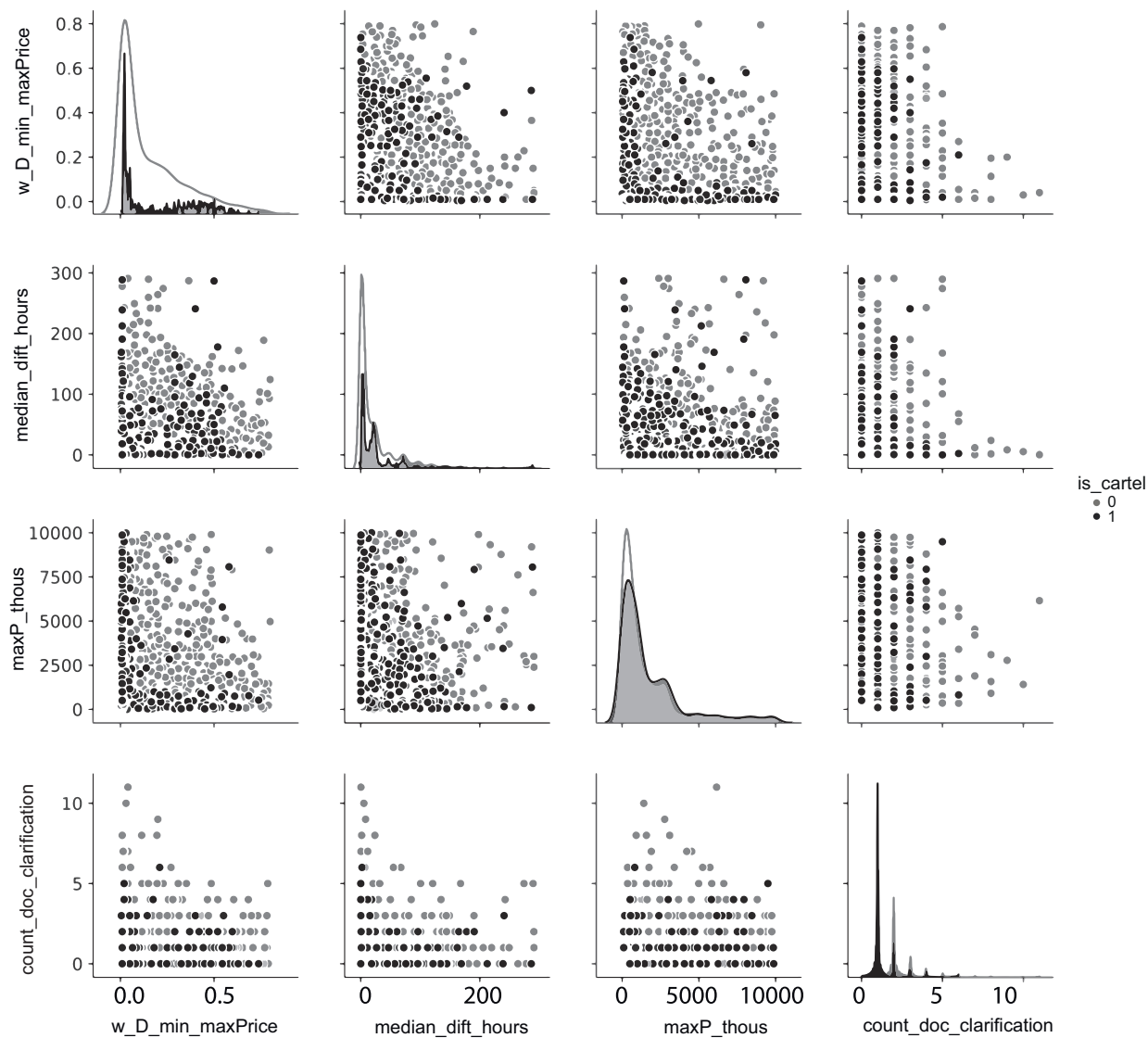


Рис. 14. Зависимости между всеми парами переменных. 1 часть

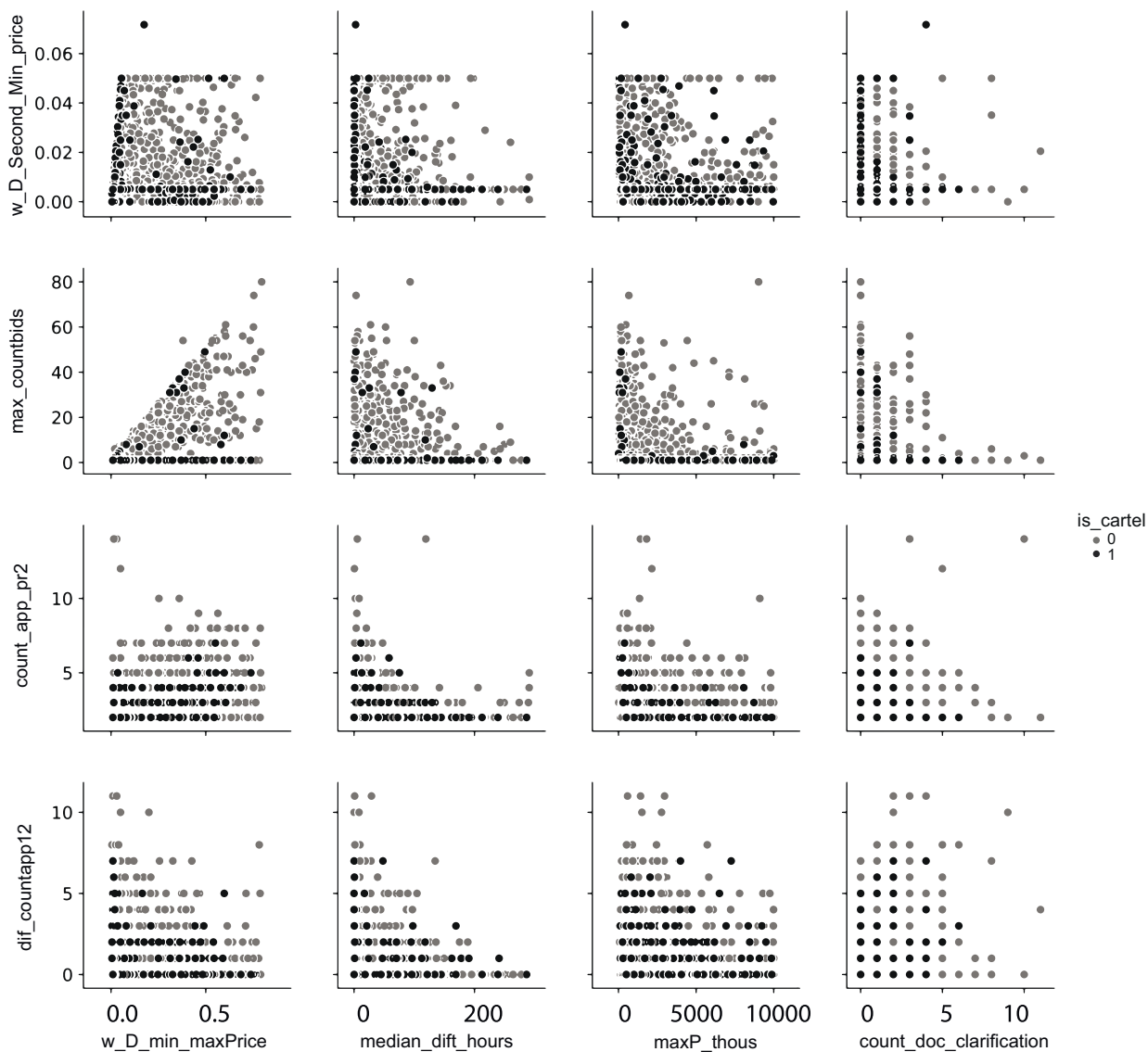


Рис. 15. Зависимости между всеми парами переменных. 2 часть

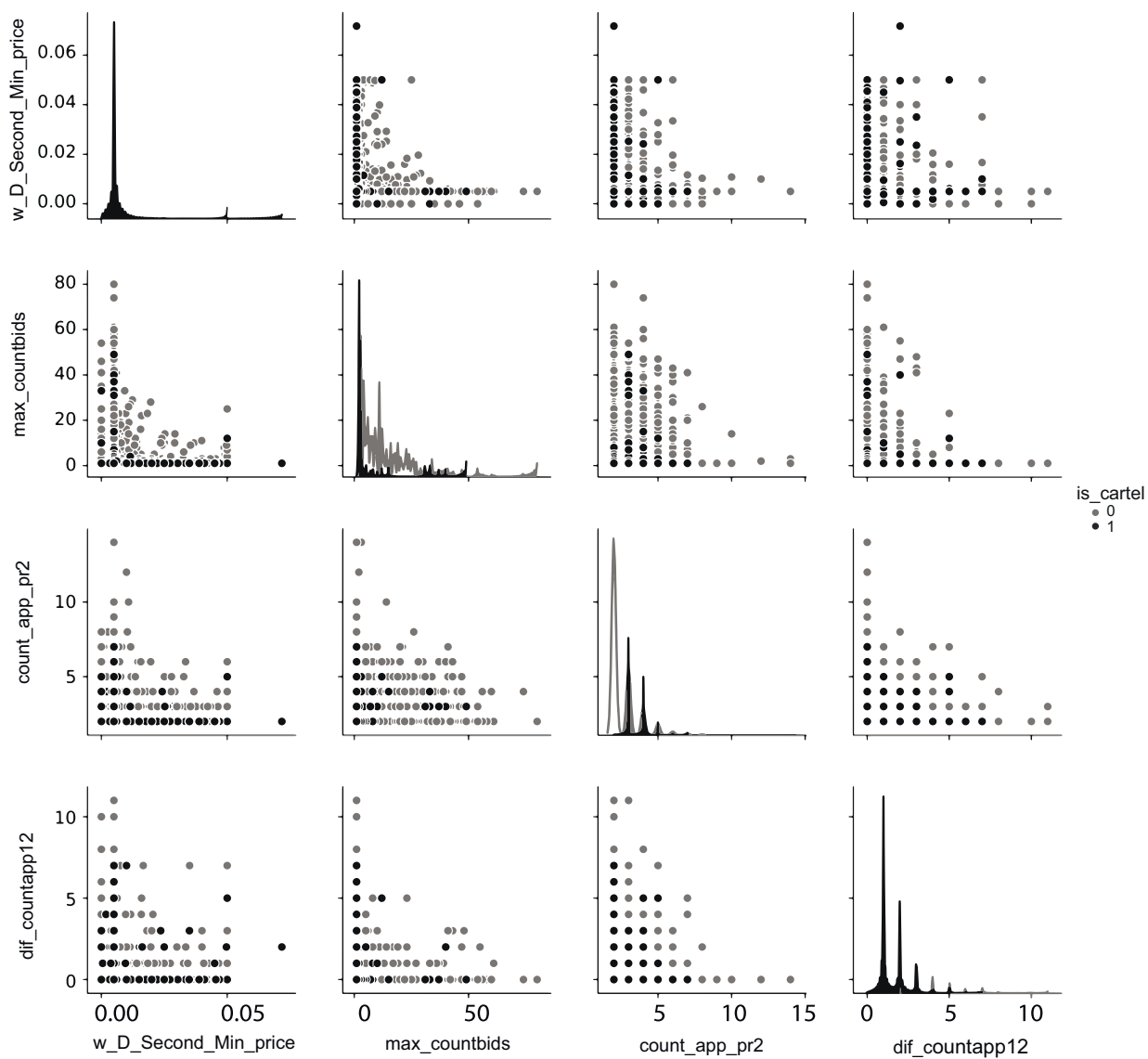


Рис. 16. Зависимости между всеми парами переменных. 3 часть

Приложение 4

Логистическая функция. Подбор параметра регуляризации

Для построения модели необходимо выбрать уровень регуляризации C . При увеличении C регуляризация ослабляется, и классификатор начинает учитывать более сложные зависимости (т.е. при $C \uparrow \Rightarrow$ сложность модели $\uparrow$). Чем меньше C , тем меньше модель штрафует за ошибки классификации, и при очень маленьких C будет проблема недообучения. Наоборот, при увеличении C в какой-то момент модель переобучится, т.е. будет слишком точно подгоняться под параметры отдельных наблюдений в обучающей выборке. Поэтому важно найти оптимальный уровень C . Для этого используем кривые правильности при обучении (Training curve) и валидации (проверке) (Validation curve). Мы применили кросс-валидацию по 10 блокам и построили несколько моделей с разными C . Результаты показывают, что наиболее оптимальным значением C является $C = 10^{-1}$ (рис. 17).

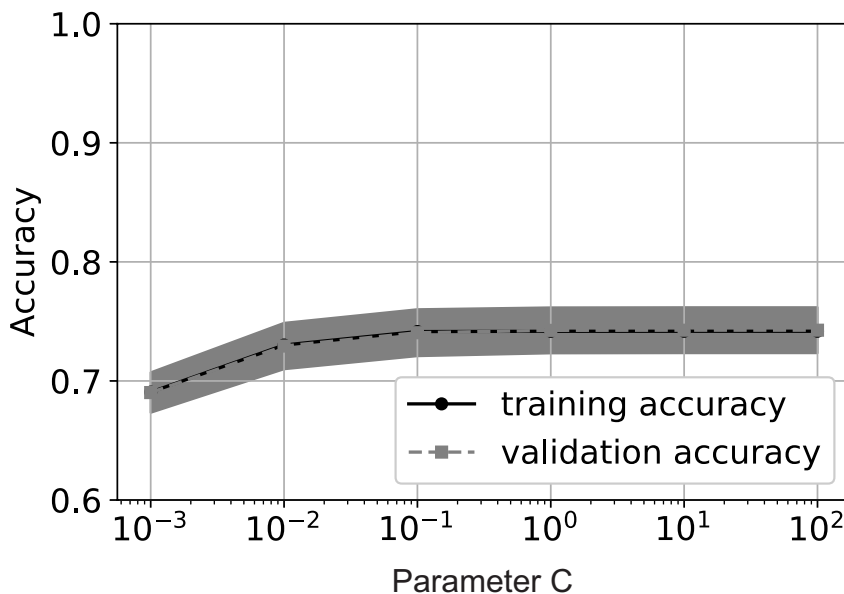


Рис. 17. Кривая проверки и кривая обучения. Логистическая регрессия

Метод k ближайших соседей (knn-модель). Определение параметра k

Для подбора параметра k используем кросс-валидацию.

Если посчитать правильность предсказаний моделей, построенных по нескольким блокам кросс-валидации, то среднее значение от полученных показателей (CV_accuracy) больше будет соответствовать тому, как будет вести себя модель на случайной выборке электронных аукционов. На рисунке ниже (Рис. 18) представлен график CV_accuracy, построенной по 10 блокам кросс-валидации, для диапазона k от 1 до 20. Для расчета k используем функцию GridSearchCV, которая n число раз обучит нашу модель для каждого значения k из некоторого заданного диапазона. Далее на основе результатов GridSearchCV выберем такой k , чтобы точность модели была максимальной. GridSearchCV, построенная по 10 блокам кросс-валидации и для k [1,20], показывает, что оптимальным является $k = 16$.

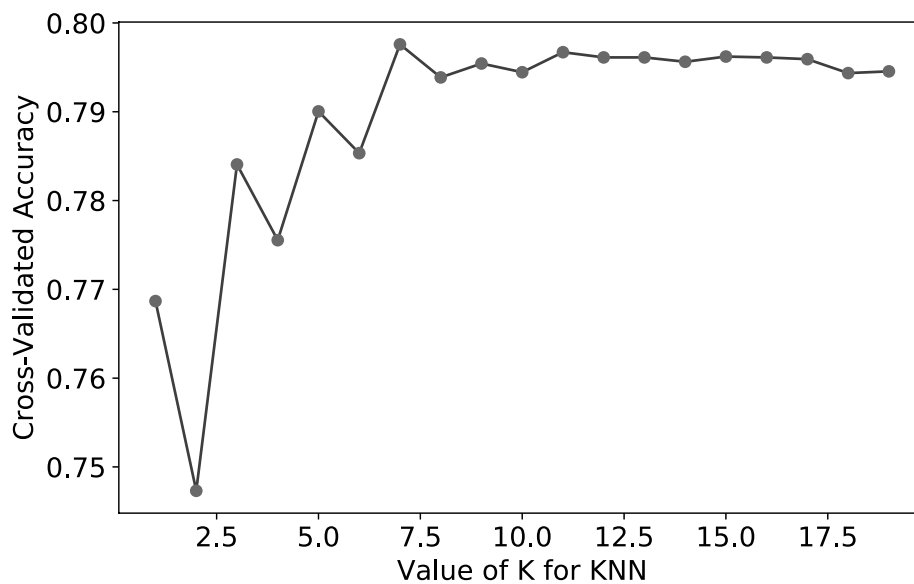


Рис. 18. График правильности классификации knn -модели в зависимости от значения k
Источник: рассчитано авторами.