Федеральное государственное автономное образовательное учреждение
высшего образования «Национальный исследовательский университет
«Высшая школа экономики»

Факультет компьютерных наук
Основная образовательная программа
Прикладная математика и информатика

ГРУППОВАЯ КУРСОВАЯ РАБОТА

Программный проект на тему

"Предиктивный анализ производительности"

Выполнили
студент группы 192, 3 курса, Тряпицын Александр Михайлович
студент группы 198, 3 курса, Гладких Роман Евгеньевич

Руководитель КР:
Руководитель направления в ПАО Ростелеком
Летуновская Елена Михайловна

Москва 2022

1 Аннотация

Предиктивный анализ производительности - задача выявления основных факторов влияющих на ключевые показатели компании и их предсказание на основе этих факторов. Данная задача зачастую решается для компаний, разбитых на независимые крупные филиалы, решающие одну и ту же задачу клиента компании и имеющие каждый различных по сферам деятельности сотрудников, поскольку позволяет выявлять зависимости между результатами работы филиалов и их внутренним устройством. Это позволяет компаниям искать точки роста отстающих филиалов, оптимизировать расходы компании на персонал, моделировать стратегию развития компании при ее расширении и повышать конкурентоспособность.

В данной работе мы сфокусируемся на анализе динамики финансовых показателей филиалов компании ПАО Ростелеком в зависимости от исторических данных о результативности этих филиалов и сведений о сотрудниках и их различных трудовых характеристик.

Predictive performance analysis is the task of identifying the main factors affecting the company's key indicators and predicting them based on these factors. This task is often solved for companies that are divided into independent large branches that solve the same task of the company's client and each have employees who are different in their fields of activity, since it allows you to identify dependencies between the results of the work of branches and their internal structure. This allows companies to look for growth points of lagging branches, optimize the company's personnel costs, model the company's development strategy during its expansion and increase competitiveness.

In this paper, we will focus on analyzing the dynamics of the financial indicators of the branches of PJSC Rostelecom, depending on historical data on the performance of these branches and information about employees and their labor characteristics.

Содержание

2 Терминология

1 B2C (Business To Consumer) - модель предоставления услуг частным лицам

2 B2B (Business To Business) - модель предоставления услуг частным компаниям

3 B2G (Business To Government) - модель предоставления услуг государственным органам и компаниям

4 Временной ряд - последовательность чисел, порядок которых описывается некоторым временем, соответствующим каждому числу, описывающая изменение некоторой величины с течением времени

5 SOTA (State Of The Art) - лучший к текущему моменту. Применяется по отношению к самым лучшим и современным в своем классе методам

6 ML (Machine Learning) - термин для обозначения методов, относящихся к машинному обучению и искусственному интеллекту

7 Бейзлайн (от англ. baseline) - термин для обозначения качества референсного решения некоторой задачи

3 Введение

3.1 Обзор исследуемой области

Сегодня множество больших компаний и корпораций вкладывают ресурсы в развивитие нового направления внутренней аналитики. В рамках этого направления компаниями исследуются и тестируются различные гипотезы, направленные как на повышение производительности отдельных групп сотрудников за счет выявления возможностей влияния копрорации на их рабочий процесс, так и на оптимизацию интегральных показателей компании, за счет нахождение точек роста во внутреннем устройстве и корпоративной культуре.

Компании и целые страны по-разному используют результаты внутренней аналитики. Некоторые, экспериментируют с длительностью рабочей недели[1], некоторые - сокращают неэффективных по результатам исследований данных сотрудников[2]. Таким образом, современные компании предпринимают попытки анализа рабочих процессов и расширяют сферу влияния внутренней аналитики на принимаемые решения.

С развитием направления внутренней аналитики в индустрии происходит развитие методов и подходов в этой области. Развитие методологии внутренней аналитики позволяет получать более обощенные результаты, применимые не только для конкретных компаний, но и для более широкого спектра похожих организаций. Как правило, если результаты полученные в одной компании не применимы к другой, вследствие, например, различия в сферах деятельности или региональных политик, они дают интуицию относительно возможных гипотез, которые следует проверить. С расширением методолгии внутренней аналитики можно связать появление в последние годы большого числа статей и результатов в области анализа работы корпораций.

К наиболее активно развивающимся методам внутренней аналитики мож-

[1] https://www.washingtonpost.com/business/2021/07/06/iceland-four-day-work-week/

[2] https://www.thetimes.co.uk/article/computer-says-clear-your-desk-artificial-intelligence-wfh-covid-hc7wvs3s3

но отнести hr-аналитику, использующие методы статистики и машинного обучения для тестирования гипотез о зависимости показателей компаний от трех основных факторов:

1 Устройства этапа найма новых сотрудников

2 Распределения сотрудников по различным характеристикам

3 Рабочего процесса сотрудников

Примером значимых результатов в области hr-аналитики можно назвать развитие методов анализа влияния на компанию различий в условиях труда сотрудников, отличающихся по этническим или гендерным характеристикам, а также анализ подходов к разрешению этих различий (Astley and Cherkashyna (2021)).

В качестве другого примера можно привести результат исследования эффективности подходов машинного обучения к решению проблемы распределения задач и компенсаций сотрудников компаний (Robert et al. (2020)). Одним из важнейших направлений исследования был анализ способности методов машинного обучения к справедливому распределению компенсаций.

К наиболее практическим результатам hr-аналитики можно отнести исследование текучести кадров организаций и способов удержвания сотрудников, где выявлялись основные не экономические факторы, влияющие на решение сотрудника об уходе из компании (Ribes et al. (2017)).

В рамках данной работы будет проведен предиктивный анализ производительности российской компании ПАО Ростелеком (далее - Ростелеком). Компания является одним из крупнейших в России провайдером цифровых услуг в рамках каждой из моделей B2C, B2B и B2G.

Ростелеком предоставляет свои услуги на большей части территории России и разделен на соответствующие региональные филиалы. Всего в компании насчитывается более 50 филиалов, расположенных в различных субъектах государства. Каждый филиал кроме прочего занимается подключением новых клиентов и поддержкой текущих.

Разделение компании Ростелеком на условно независимые неконкурирующие филиалы, выполняющие схожие функции, дает возможность эффективно применять к компании методы hr-аналитики как для сравнительного межфилиального анализа и выявление основных предпосылок к достижению высоких месячных и годовых результатов работы, так и для предсказания ключевых показателей компании для их использования в моделировании стратегии компании. Ростелеком, как и любому публичному акционерному обществу, традиционно важно иметь наиболее точные прогнозы по ключевым показателям, для улучшения биржевой стабильности.

4 Обзор литературы

4.1 Обзор предметной области

Как правило задачи hr-аналитики решаются внутри частных и государственных компаний и публикация програмных решения зачастую не является частью статей о результатах исследований в связи с политиками неразглашений этих компаний, как и количественное измерение этих результатов. Из таких публикаций мы рассмотрим интуицию методов и идей, и будем развивать ее в рамках целеполагания текущей задачи.

Тем не менее, так как задачи hr-аналитики сводятся к применению традиционных методов машинного обучения и статистики, в качестве методологической литературы будем рассматривать современные статьи по машинному обучению, применимые к нашим задачам.

При математическом моделировании проблем hr-аналитики можно выявить несколько основных базовых задач:

1 Задача кластеризации для разделения филиалов по их характеристикам и результатам

2 Задача интерпретации основных факторов предсказательных моделей для выявления ключевых характеристик филиалов

3 Задача предсказания продолжений временных рядов для прогнозирования динамики целевых показателей

В данной работе будет рассмотрена по большей части задача, связанная с установлением закономерностей во временном ряде и продолжением его вперед для формирования ожидаемых значений ключевых переменных.

4.2 Обзор современных подходов

В настоящий момент времени не существует подхода, однозначно признанного лучшим в области в задаче прогнозирования временных рядов. Лучший

подход определяется путем проведения ряда эксперементов над моделями, описывающими временной ряд. Рассмотрим несколько путей, являющихся основополагающими в соврменной науке.

1 ETS (Error, Trend, Seasonal) Holt (1957) - классическое семейство моделей для прогнозирования временных рядов, представленное более 50 лет назад. Основывается на разложении временного ряда на 3 компоненты, где у каждой компоненты может быть от 3 до 4 значений (“N” = нет, “A” = аддитивный, “M” = мультипликативный, “Z” = выбирается автоматически):

1.1. Тип ошибки (“A”, “M”, “Z”)

1.2. Тип тренда (“N”, ”A”, ”M”, “Z”)

1.3. Тип сезона (“N”, ”A”, ”M”, “Z”)

После разложения на компоненты идет моделирование каждой из них. В настоящие дни ETS все еще считается достаточно хорошей моделью в некоторых ситуациях. Goodwin (2010)

2 ARIMA (Auto Regressive Integrated Moving Average) Asteriou and Hall (2011) - семейство моделей описания временных рядов, использующих в своей основе принципы моделирования стационарных процессов суммами белых шумов, построенные на авторегрессионном анализе. Данная модель является комбинацияей двух более простых моделей, с добавлением дифференциации:

2.1. Авторегрессионная модель (AR)

Идея: предположим, что предыдущие значения влияют на нынешние. Таким образом мы можем построить линейную регресиию, для предсказания текущих значений черех прошлые. Порядок AR модели равен количеству временных отметок за которые есть данные.

2.2. Модель скользящей средней (MA)

Идея: предположим, что предыдущие условия ошибок влияют на нынешнее значение переменной.

Дифференцирование в модели ARIMA используется для преобразования временного ряда в стационарный, путём вычитания текущее значение из предыдущего. Данная модель имеем 3 параметра (p, d, q):

- p: порядок AR модели
- d: количество несезонных различий (порядок дифференцирования)
- q: порядок MA модели

SOTA модель прогнозирования такого типа имеет название Prophet Letham and S.J. (2017) и способна учитывать точки изменения тренда ряда.

3 Методы машинного обучения Balkin and Ord (2000) - наиболее современное семейство методов, показывающее лучшие результаты. Наиболее популярным является ипользование рекуррентных нейронных сетей Sherstinsky (2018). Проблема метода заключается в невозможности интерпретировать причинно следственные-связи. Однако, данное семейтство методов позволяет лучше других учитывать контекст точек временного ряда. Часто, методы машинного обучения используют для прогнозирования остатков поверх классических методов Hardik Goel (2017).

5 Постановка задачи

Компанией Ростелеком были предоставлены данные о месячной доходности[3] 61 филиал на территории Российской Федерации за 2019-2021 г. и учетные книги сотрудников за 2019-2020 год.

Среди филиалов есть как небольшие городские филиалы, так и крупные региональные центры. Доходности последних могут в значительной степени различаться.

В учетных книгах сотрудников имеются обезличенные данные о стаже сотрудника, его должности, поле, возрасте, количестве отрабаотанных дней в месяцах и грейде. Чем больше грейд сотрудника, тем он более хорошим специалистом он считается в компании.

В этой работе будет рассмотрено множество алгоритмов прогнозирования динамики ключевых показателей на основе данных по персоналу с применением методов статистики и машинного обучения и выбран лучший.

В нашей задаче целевой показатель - месячная доходность филиала.

[3]Из-за ограничений на распространиние данных, в работе будут продемонстрированы модифицированные показатели

6 Анализ данных

Формально мы имеем 61 временной ряд, соответствующих разным филиалам, по 36 наблюдений в каждом за один и тот же промежуток времени длительностью 3 года, где каждое наблюдение отражает доходность филиала за месяц. Иллюстрация финансовых данных приведена на рисунке 6.1. Наблюдения за первые 2 года сопровождаются данными о персонале и будут использованы для обучения предсказательных моделей, а последний год (2021) - для их тестирования.

Филиал	алтайский филиал	амурский филиал	архангельский филиал	астраханский филиал	белгородский филиал	бурятский филиал	волгоградский филиал
Date							
2019-01	3937.117871	2113.044960	2853.736310	1254.366592	2808.848539	3475.777877	2994.418594
2019-02	4106.833764	2050.159387	2850.674599	1333.782801	2656.541492	3502.798131	2980.140202
2019-03	4059.795418	1931.034934	2935.491051	1964.252253	2682.795968	3568.251474	2647.504803
2019-04	3938.053286	1989.886445	3054.751214	1346.579877	2651.333930	4036.351257	3042.545366
2019-05	4095.868310	2034.197688	2919.342064	1353.759978	2638.473679	3564.733199	3032.980294

Рис. 6.1: Пример устройства данных

В этом разделе мы сфокусируемся на поиске закономерностей в данных для их дальнейшего использования и рассмотрим следующие аспекты:

1 Поведенческий анализ ряда: поиск тренда и сезонности

2 Анализ автокорреляций и частных автокорреляций: поиск сезонности и цикличности

3 Анализ взаимной корреляции различных рядов

Здесь и далее для упрощения анализа сразу многих временных рядов введем усредненный ряд, значения которого получаются как среднее значений

целевого показателя по филиалам. Обоснованность и некоторую достаточность анализа такого ряда покажем ниже в разделе об анализе взаимных корреляций.

6.1 Поведенческий анализ ряда

На графике 6.2 изображено значение целевой переменной. Видно, что ряд обладает некоторыми свойствами:

1 В течение месяцев одного года выручка растет

2 В декабре выручка принимает пиковое значение

3 В январе выручка резко падает по сравнению с декабрем

В результате консультаций с бизнес-партнерами из Ростелекома получилось в значительной степени объяснить полученные наблюдения следующими двумя причинами:

1 В течение года бизнес расширяется, а деньги обесцениваются. Как результат - выручка компании постепенно растет

2 Многие менеджеры склонны заключать сделки в декабре, а не оставлять их на январь. Это бюрократически удобнее для компании и может влиять на ежегодные компенсации основных задействованных в сделке сотрдников.

На графике 6.2 также изображены результаты STL разложения (Cleveland et al. (2019)) усредненного ряда, с параметром сезонности в 12 месяцев. По результат можно сделать 3 вывода:

1 Тренд меняет свое направление. Это можно связать с пандемией, начавшейся в 2021 году

2 Сезонность практически отсутствует в месяцы, отличные от декабря. Мы наблюдали это и визуально.

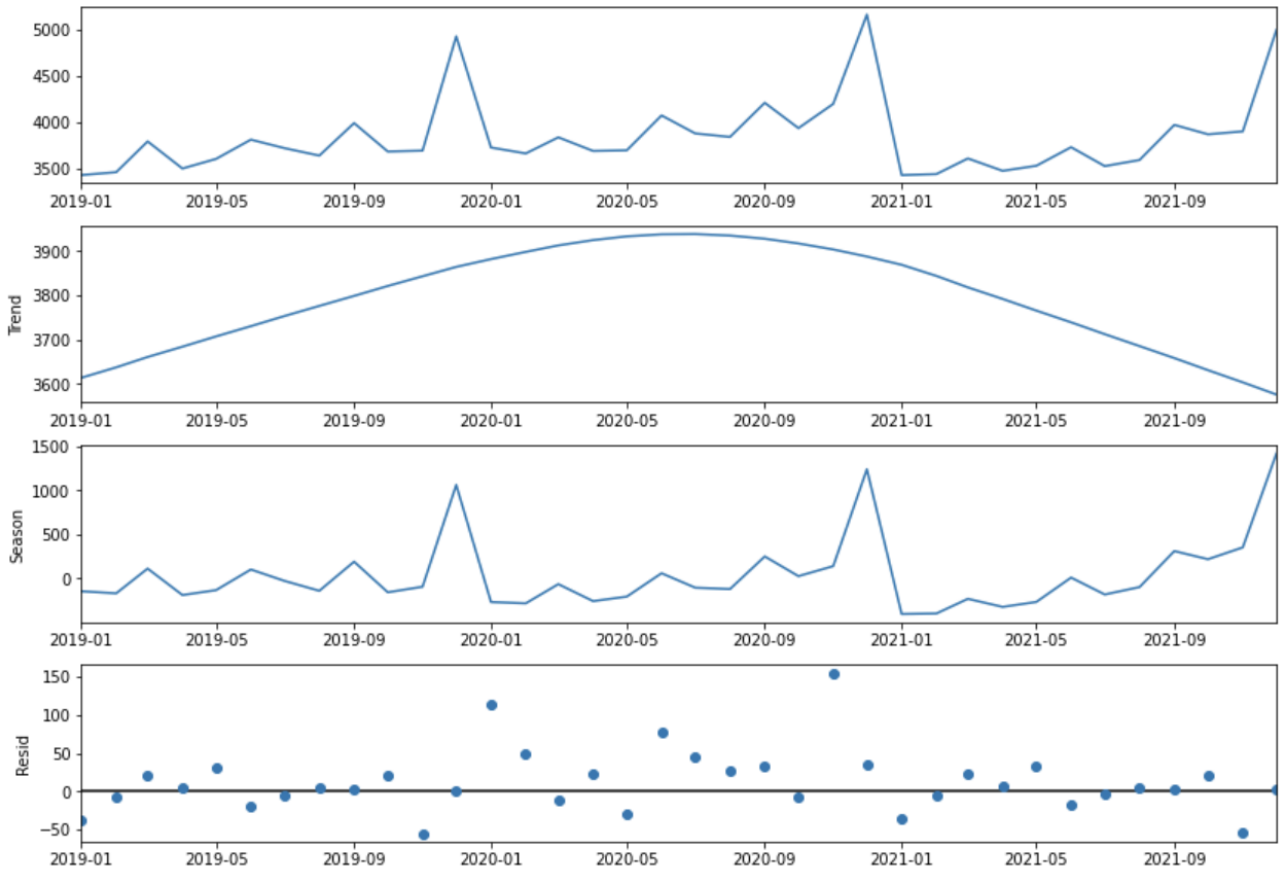


Рис. 6.2: STL разложение усредненного ряда. Сверху вниз: значение целевой переменной, тренд, сезонность, остаток

3 Тренд и сезонность хорошо описывают данные, показывания малый уровень ошибки на уровне 2%, а значит у референсного решения с помощью наивного предсказания также можно ожидать хорошее качество с ошибкой порядка $2\sqrt{61} = 15\%$ в предположении независимости рядов, то есть около 390.

6.2 Анализ автокорреляций и частных автокорреляций

Автокорреляции - корреляции между членами одного и той же последовательности - отражают силу и статистическую значимость линейный зависимостей членов последовательностей, взятых со сдвигом. Нахождение значимых автокорреляций позволит получить предпосылылки относительно параметров семейства ARIMA моделей прогнозирования временных рядов.

Можем видеть высокую корреляцию ряда с собой, сдвинутого на 12. Это является предпосылкой для использования лагов двенадцатого порядка в

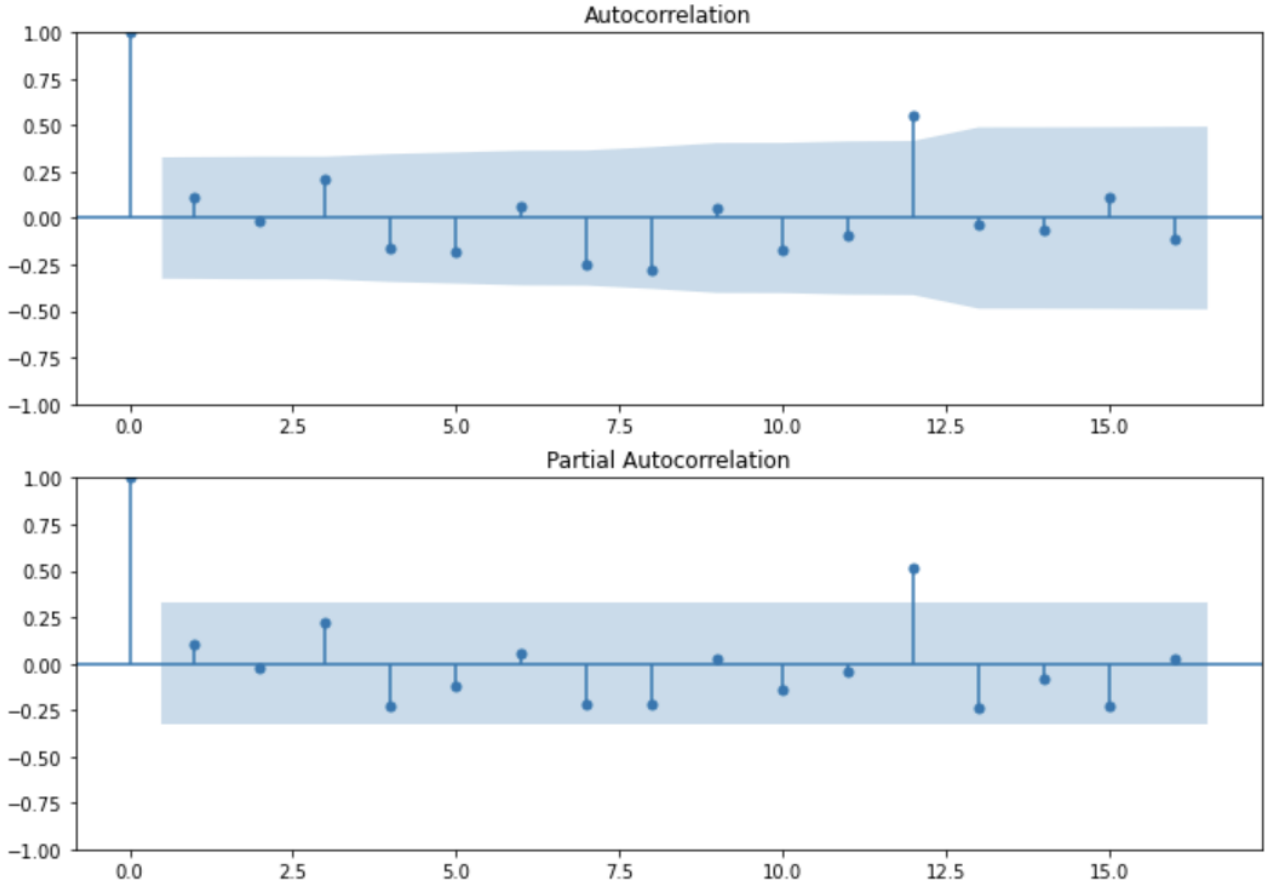


Рис. 6.3: Автокорреляции (сверху) и частные автокорреляции (снизу) усредненного ряда. По Ox - лаг ряда.

предсказательных моделях.

6.3 Анализ взаимной корреляции различных рядов

Вследствие ограниченности наших данных во времени, нам нужно извлекать из них как можно больше информации. Одним из способов учесть дополнительные факторы о временном ряде будет учет изменений сопутствующих рядов. Этот признак должен хорошо себя показать, если ряды в значительной степени коррелируют.

Согласно распределению попарных корреляций рядов, соответствующих различным филиалам, на рисунке 6.4 значительная положительная корреляция во многом имеет место быть. Этот эффект вполне объясним: все филиалы Ростелекома находятся на территории России и на них действуют одинаковые внещние факторы, как политического, так и экономического рода.

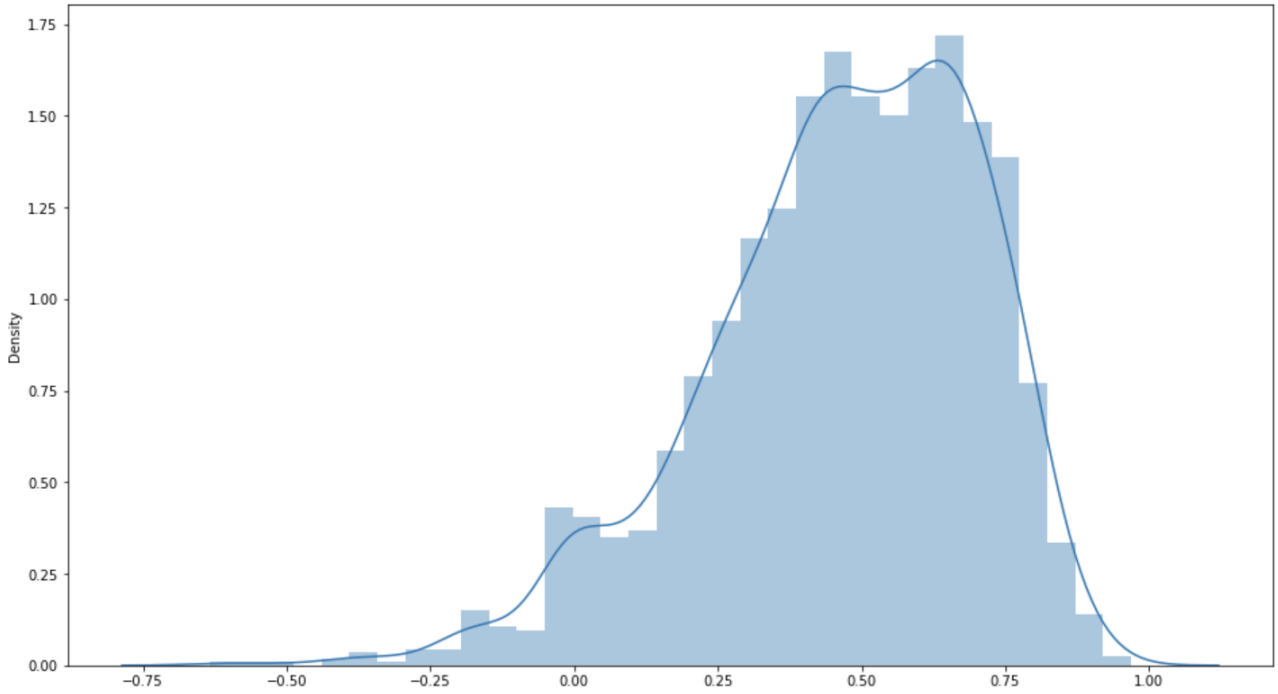


Рис. 6.4: Распределение попарных корреляций рядов, соответствующих различным филиалам Ростелекома. По Ox - значение корреляции.

Кроме того, высокая степень взаимной скоррелированности рядов позволяет нам проводить аналитику усредненного ряда, так как его корреляция с каждым рядом, соответствующим отдельному филиалу, будет также высока.

7 Отбор признаков

Зафиксируем для модели 10 лучших признаков и посмотрим поподробнее на самые интересные.

7.1 Кумулятивная инфляция

С помощью таблицы плотности распределение корреляций инфляции и выручки филиалов 7.1 мы можем заметить, что данный ряд сильно удален от нуля, а также очень плотно распределен. Что показывает большое влияние данного признака на целевую переменную

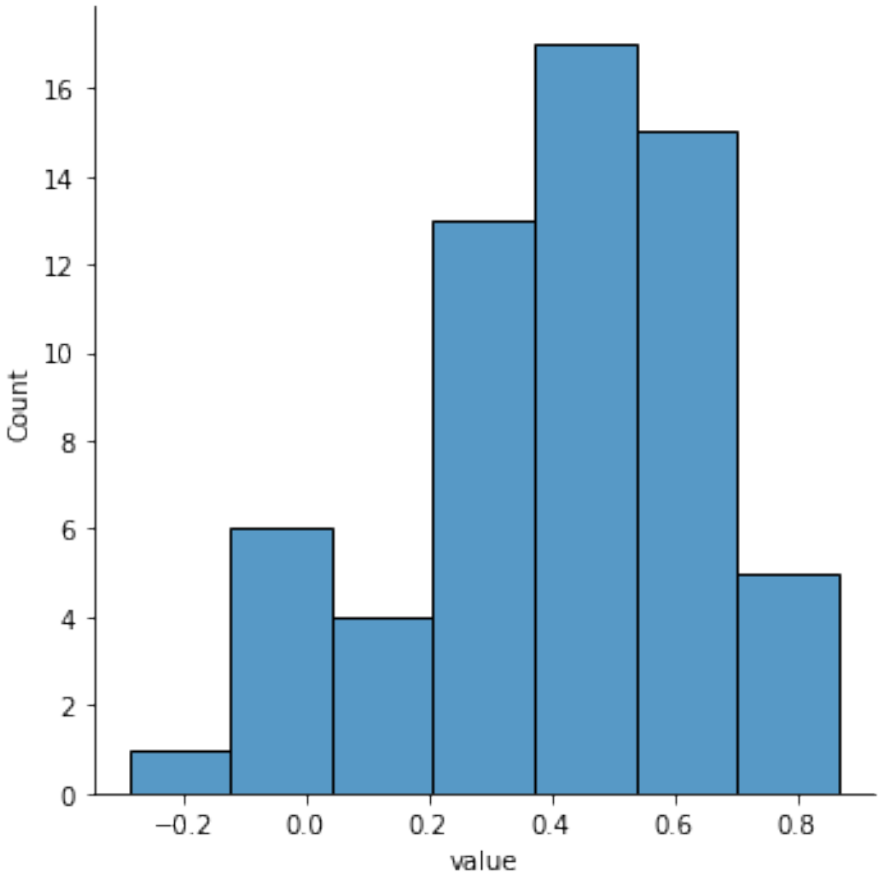


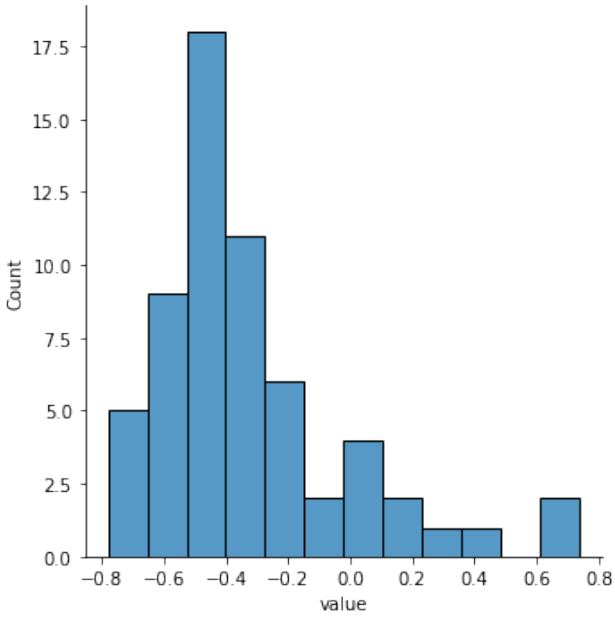
Рис. 7.1: Распределение корреляций инфляции и выручки филиалов

7.2 Суммарное количество рабочих ставок

С помощью таблицы плотности распределения корреляций и диаграммы рассеяния 7.2, заметим что несмотря на то что большинство членов данного ряда меньше 0, мы видим что данный ряд далёк от нуля, а также что он плотно распределён. Немаловажно, что его диаграма рассеяния имеет формулу узкого треугольника с небольши количеством выбросов. А мы знаем, что чем ближе ряд к линейной функции, тем ценнее данный признак. По этому показателю, данный признак является одним из лучших.

7.3 Курсы акций Ростелеком и МТС

По таблицам плотности распределения корреляций курсов акций и выручки филиалов 7.3, видно что данные распределения далеки от нуля и плотно распределены, это также можно заметить на общей таблице корреляции между признаком и выручкой 8.1. Курс акций МТС, имеет самое плотное



Распределение корреляций

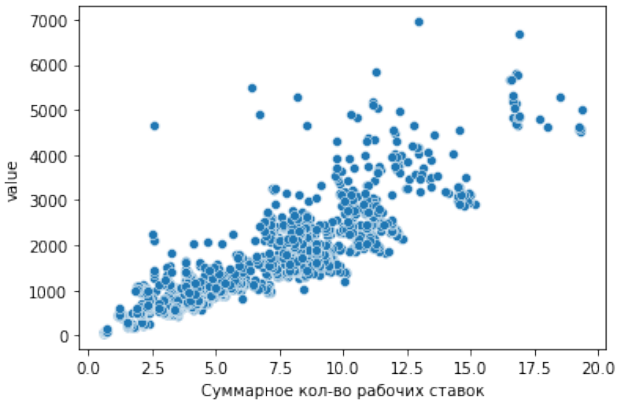
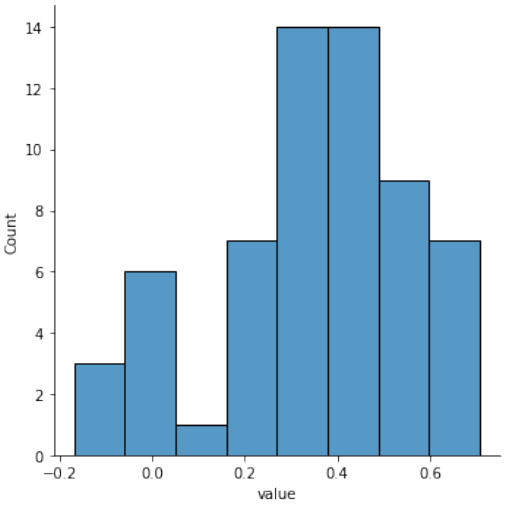


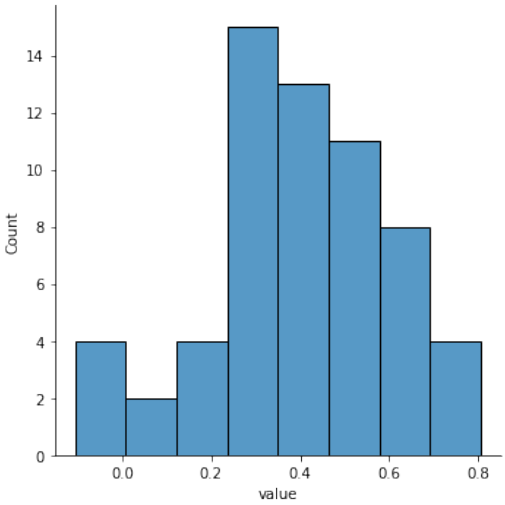
Диаграмма рассеяния

Рис. 7.2: Распределение корреляций и диаграмма рассеяния кол-ва рабочих ставок и выручки филиалов.

распределение среди всех, это видно по наименьшему модулю разницы медианы и среднего расстояния до нуля, а также заметим, что его расстояние между персцентилями минимум на треть меньше чем у остальных признаков входящих в 10 лучших. Курс акций Ростелеком, хоть и похуже чем МТС, но не смотря на это тоже имеет достаточно плотное распределение.



Ростелеком



МТС

Рис. 7.3: Распределение корреляций курса акции и выручки филиалов

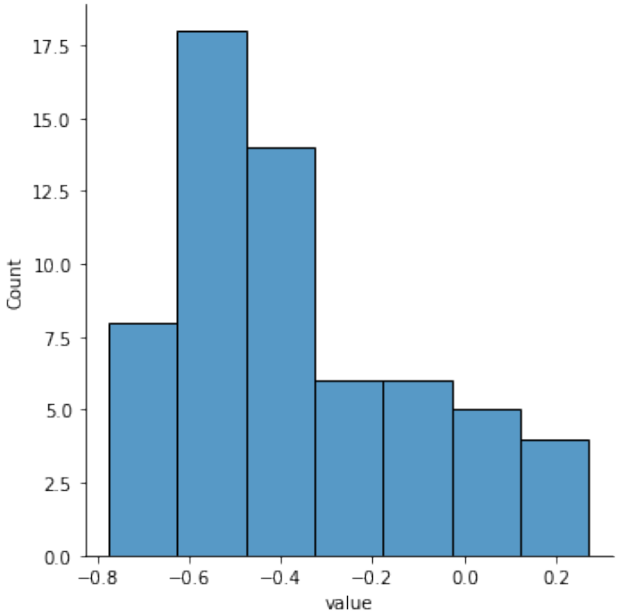
7.4 Количество женщин

Еще одним признаком, который вошёл в тройку самых важных стало количество женщин, работающих в филиале. Мы не нашли однозначной связи этого показателя с результирующим, но имеем 2 гипотезы по этому вопросу.

1 Количество женщин, как и количество сотрудников, влияет на размер суммарной заработной платы, уплачиваемой компанией, а значит к снижению прибыли при неизменной эффективности

2 Из-за несимметричности распределния мужчин и женщин по профффесиям, могла сложиться ситуация, когда в филиалах, получающих бОльшую прибыль, была бОльшая необходимость в тех профессиях, где преоблажают женщины

На таблице плотности распределения корреляций и диаграмме рассеяния 7.4 видно что данный ряд имеет очень плотное распредление, а также очень узкий треугольник рассеяния с малым количеством выбросов.



Распределение корреляций

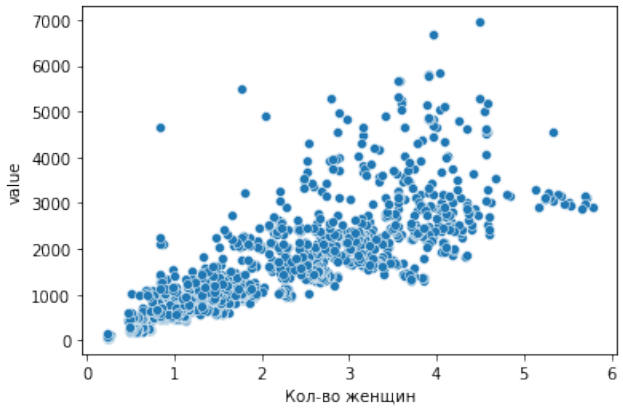


Диаграмма рассеяния

Рис. 7.4: Распределение корреляций и диаграмма рассеяния кол-ва женщин и выручки филиалов.

8 Отбор признаков

Для эффективного использования широкого спектра моделей требуется отобрать самые важные и полезные признаки.

8.1 Метод отбора

Для отбора будем пользоваться данными по трем графикам:

1 Считаем попарную корреляцию между признаком и целевым значением для каждого филиала, по итогу получаем 61 значение $\in [-1, 1]$.

Интуиция: мы смотрим на разницу между 61 временным рядом для понимания влияния изменения значения признака на них

2 Считаем попарную корреляцию между признаком и целевым значением для каждого месяца, по итогу получаем 24 значения $\in [-1, 1]$.

Интуиция: мы смотрим на влияние признака на филиал в целом

3 Диаграмма рассеяния для признака и целевого значения

Интуиция: смотрим рассеяние на графике, чем сильнее график похож линейную функцию, тем больше зависимость между признаком и выручкой для каждого филиала

8.2 Интерпретация методов отбора

На таблице 8.1 результатов попарной корреляцию между признаком и выручкой для каждого филиала, мы ввели несколько метрик для измерения важности признака.

1 Среднее расстояние до 0 по модулю

Интуиция: чем больше расстояние, тем ближе средняя корреляция к 1, т.е тем сильнее корреляция между признаком и выручкой.

2 Медиана

Интуиция: смотрим на "настоящее"среднее, избавляясь от влияния выбросов.

3 90% персцентиль - 10% персцентиль

Интуиция: убираем выбросы, и смотрим на разницу между самым большим и самым маленьким элементом. То есть оцениваем плотность распределения ряда корреляций. Чем меньше - тем плотнее.

	Признаки	Среднее расстояние до 0	Медиана	90% - 10% перс.
0	Инфляция кумулятивная	0.40	0.45	0.67
1	Курс акций МТС	0.39	0.38	0.46
2	Кол-во женщин	0.36	0.44	0.64
3	Курс акций РТК	0.34	0.37	0.62
4	Процент мужчин	0.34	0.40	0.61
5	Процент женщин	0.34	0.40	0.61
6	Суммарное кол-во рабочих ставок	0.32	0.40	0.67
7	Средний грейд	0.31	0.33	0.74
8	Кол-во работников	0.30	0.35	0.83
9	Кол-во работников с высшим обр.	0.29	0.34	0.63
10	Курс доллара	0.21	0.23	0.57
11	Кол-во мужчин	0.19	0.27	0.86
12	Объём акций МТС	0.17	0.18	0.39
13	Процент работников с высшим обр.	0.09	0.20	0.83
14	Объём акций РТК	0.05	0.04	0.50
15	Средний возраст	0.01	0.01	1.07

Рис. 8.1: Попарная таблица корреляции между признаком и выручкой

9 Добавление финансовых признаков

Мы рассчитываем добавить 3 новых признака, для проверки гипотезы о том, что не только внутренние процессы, но и внешние факторы оказывают влияние на доход компании.

9.1 Курс доллара

Доллар на данный момент, является самой торгуемой валютой в мире, 87% валютных сделок проводятся с его помощью. На рисунке 9.1 видно резкое повышение курса доллара в период с декабря 2019 по март 2020. Для компании, закупающей оборудование за границей, такое резкое повышение курса доллара может сказаться на выручке. Поэтому мы проверим это.

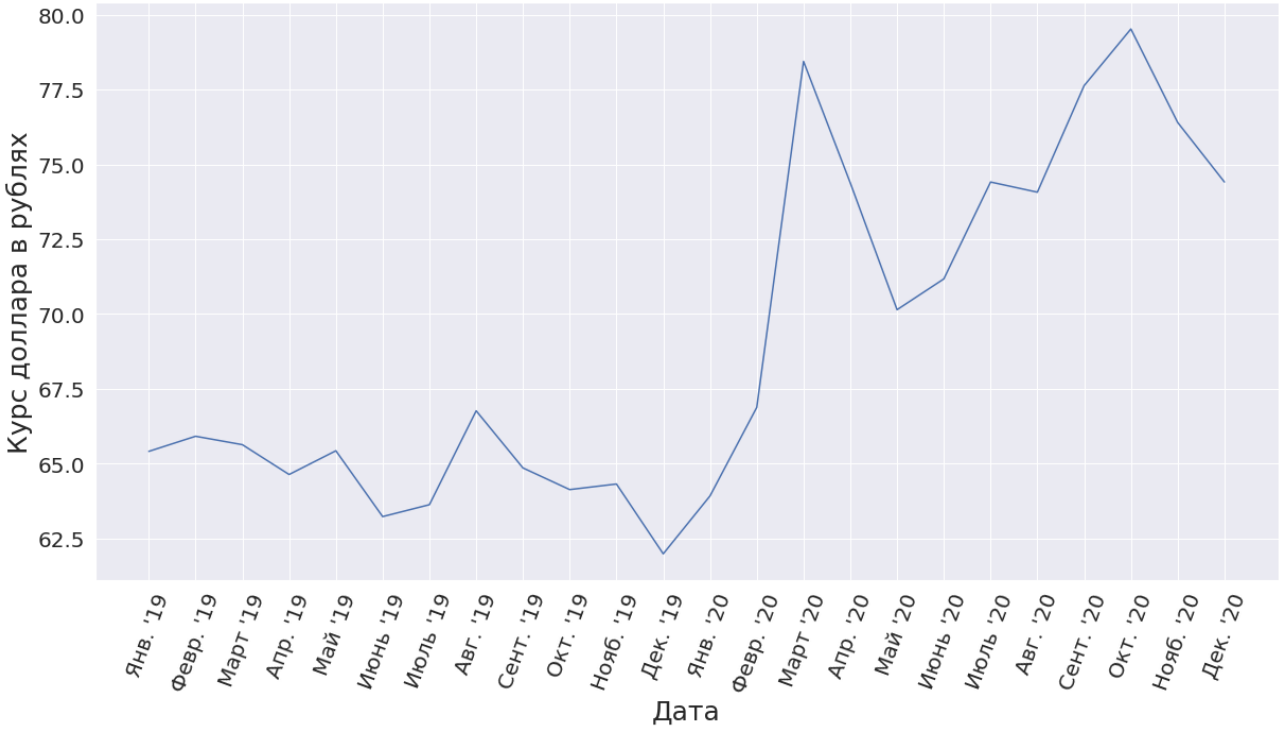


Рис. 9.1: Средний помесячный курс доллара

9.2 Кумулятивная инфляция

Инфляция неразрывно связана со ставкой Центрального Банка России. Многие компании используют кредиты для закупок оборудования. На рисунке 9.2 видны два скачка инфляции: ноябрь 2019 по март 2020 и сентябрь

2020 по декабрь 2020. Мы хотим оценить влияние роста инфляции на выручку, а также оценить 'настоящее', а не номинальное изменение выручки компании.

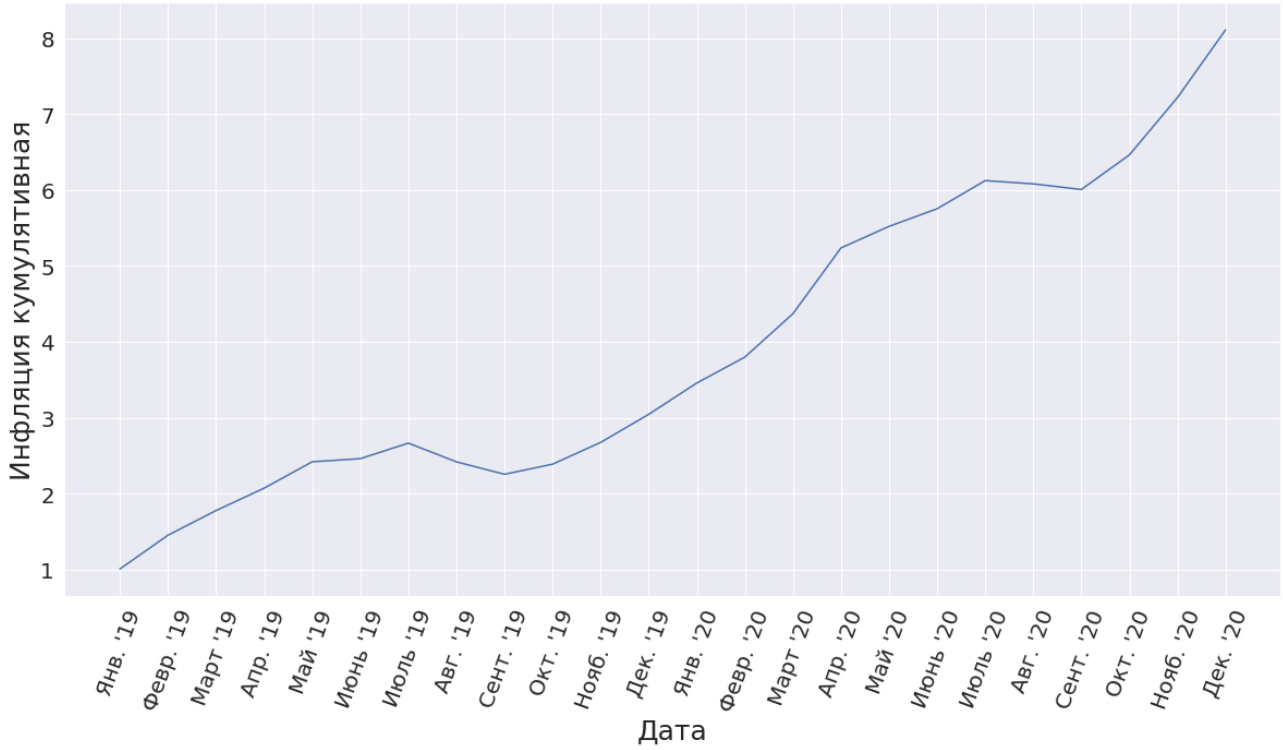


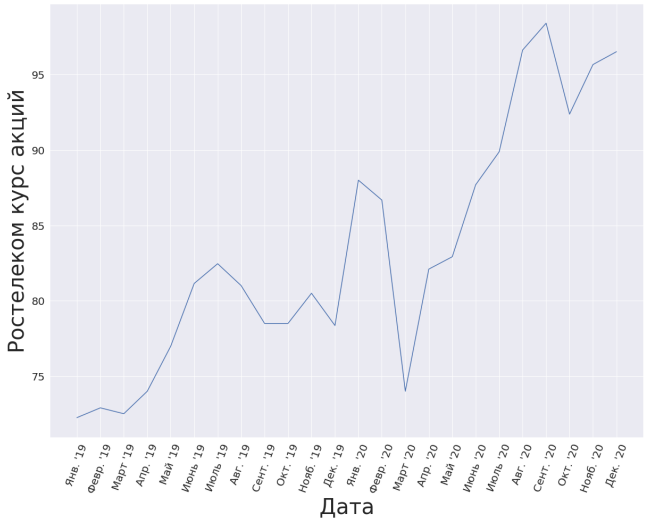
Рис. 9.2: Кумулятивная инфляция

9.3 Курс и объём продажи акций Ростелеком и МТС

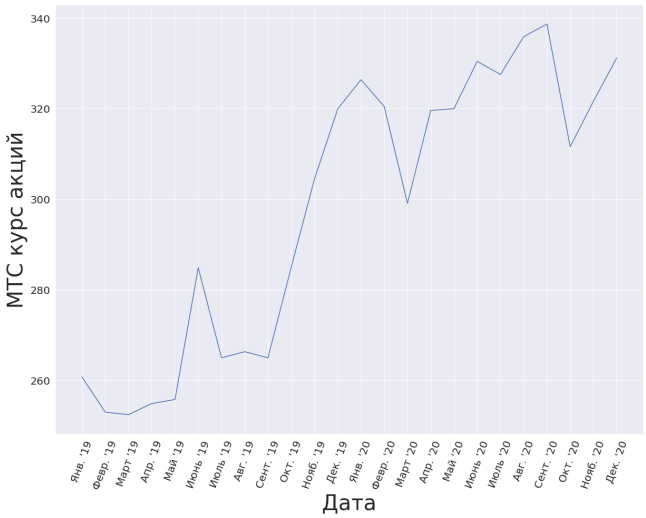
Главным конкурентом Ростелекома в сфере телекоммуникаций является МТС, эта компания занимает второе место по выручке, количеству сотрудников и длине проложеных кабелей. Поэтому мы выбрали данную компанию, как основного конкурента. Нами было решено взять цену акций двух данных компаний для отслеживания привлекательности компании. Но не только одни акции не отражают привлекательность компании, поэтому для того чтобы уравновесить различие в ценах на акции мы взяли объем продаж акций (в миллиардах рублей) для оценивания валидности акции на рынке, с помощью этого можно оценить финансовое благополучие компании.

На рисунке 9.3 можно заметить, что объем продаж резко возрос с началом локдауна в России в марте 2020, когда акции обоих гигантов резко просели

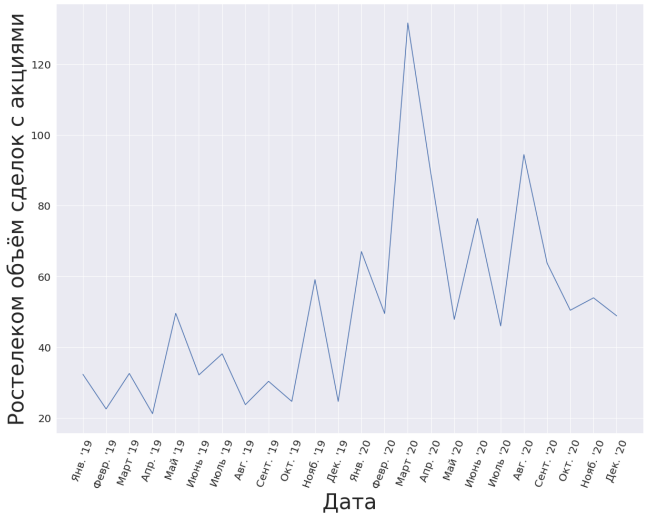
на 10-20%. Затем из-за возросшего спроса акции вернулись к прежним показателям.



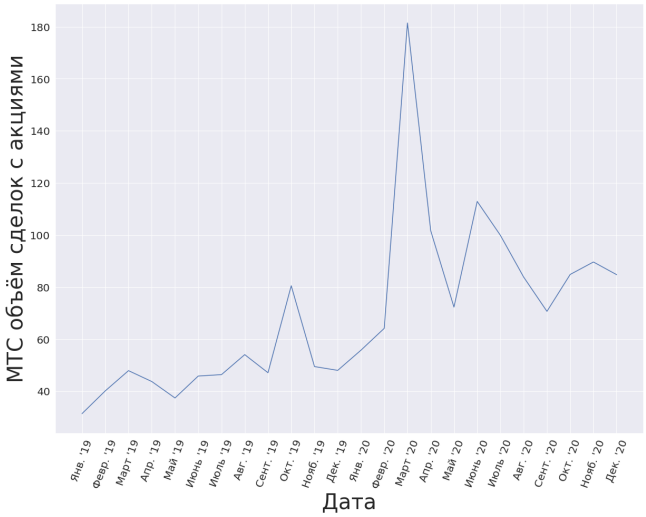
Курс акций Ростелеком



Курс акций МТС



Объём продажи акций Ростелеком



Объём продажи акций МТС

Рис. 9.3: Финансовые показатели Ростелеком и МТС

10 Проведение эксперементов

В качестве метрики качества модели прогнозирования выберем MAE (Mean Absolute Error) - среднюю величину ошибки по прогнозированной части ряда. Рассчитывать будем отдельно по каждому из рядов, соответствующих различным филиалам, а затем усреднять.

В первую очередь, проведем эксперименты с моделями, разрабатывающимися для непосредственной работы с временными рядами, а после рассмотрим модели более широкого спектра для прогнозирования данных.

10.1 Наивное предсказание

Сналача получим некоторое наивное предсказание, которое дальше будет улучшаться. Классически, в задачах связанных с временными рядами такими предсказаниями выступают либо константные, либо циклические, использующие повторение предыдущих точек с учетом сезонности.

В качестве наивного предсказания значения временного ряда за 2021 год будем использовать данные за 2020. В таком случае получим $MAE_{naive} = 340$. Это близко к 390, предсказанным ранее.

10.2 Использование ETS моделей

Семейство ETS моделей широкое и использует различные параметры для выбора конкретной модели, основные из которых определяют поведение тренда, ошибки и сезонности - мультипликативное или аддитивное. Также, например, тренд может быть дампирован, для снижения его веса в итоговом результате с течением времени. Кроме того, есть параметр регулирующий период, которому подчиняется сезонность, и в нашем случае он равен 12 месяцам для всех моделей. Параметры итоговой модели могут быть подобраны как по валидации, так и исходя из природы точек в временном ряде.

Перебрав несколько вариантов лучшей оказалась модель ETS(AAdM) с

аддитивными ошибкой, аддитивным дампированным трендом и мультипликативной сезонностью. Для ETS подхода получили $MAE_{ETS} = 930$.

Ошибка ETS модели оказалась более чем вдвое больше чем у наивной модели за счет слишком сильного полагания ETS подхода на тренд, который в самом начале тестовых данных меняется. Рассмотреть усредненный ряд, полученный по предсказаниям ETS модели, можно на рисунке 10.2

10.3 Использование ARIMA моделей

Уже было показанно, что в данных присутствует тренд, поэтому будем использовать модель SARIMA, расширяющую семейтсво моделей ARIMA для временных рядов с сезонностью.

Семейство моделей SARIMA определяется с помощью 7 основных гиперпараметров, задающих поведение ряда. Первые 3 из них являются неотрицательными целыми числами и описывают способ моделирования ряда без учета сезонности - параметр глубины авторегрессии, параметр величины окна для скользящего среднего и параметр определяющий временной ряд какого поряка дифференцирования будет моделироваться. Следующие 3 параметра определяются аналогичным образом для ряда, описывающего сезонность исходного. Последний - регулирует период, которому подчиняется сезонность. Аналогично ETS семейству моделей, параметры итоговой модели могут быть подобраны как по валидации, так и исходя из природы точек в временном ряде.

В результате перебора гиперпараметров получена модель SARIMA(0, 0, 1)(1, 1, 0, 12), обеспечивающая $MAE_{SARIMA} = 487$.

Одной из причин того, что ошибка SARIMA модели больше ошибки наивной модели, может служить авторегрессионная часть модели, интуиция позади которой - моделирование тренда. Но как мы уже показали, тренд в задаче преломляется и не может быть хорошо смоделирован без использования других данных. Рассмотреть усредненный ряд, полученный по предсказаниям

ARIMA модели, можно на рисунке 10.2

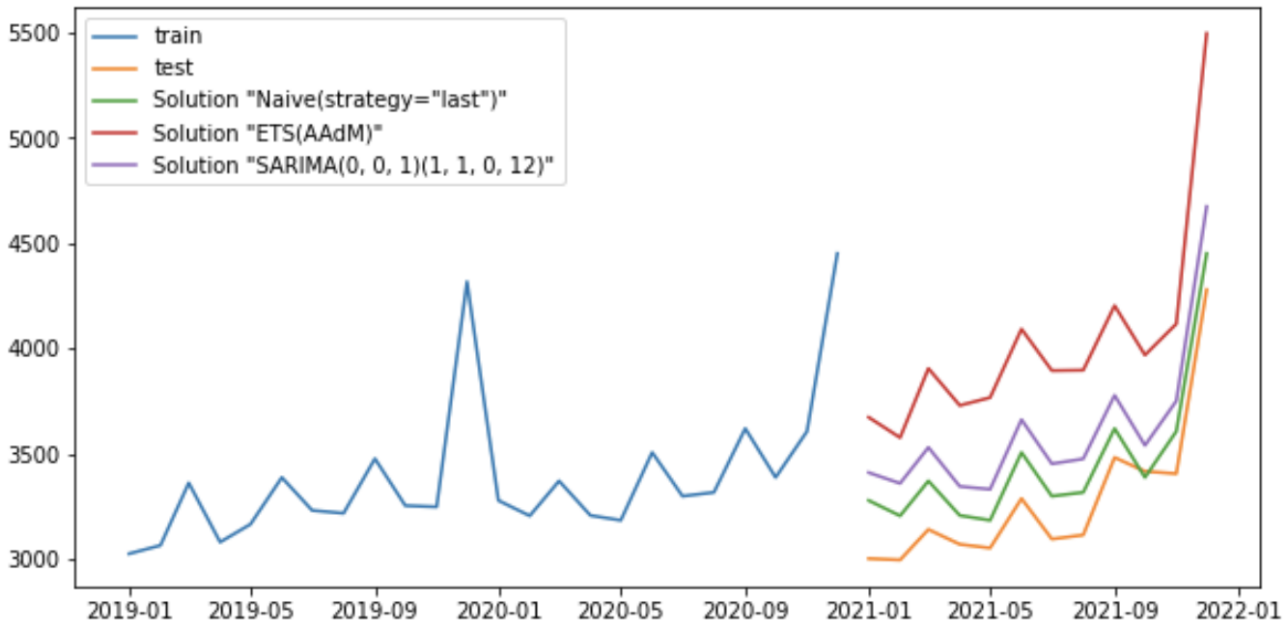


Рис. 10.1: Усредненные ряды по реальным данных и по прогнозам различных моделей, использующихся для работы непосредственно с времеными рядами

10.4 Использование VAR моделей

VAR (Vector Auto Regressive) модели - семейство авторегрессионных моделей, использующихся для предсказания сразу нескольких временных рядов. Главное отличие этой модели от нескольких AR моделей - способность учитывать взаимную информацию о рядах, что позволяет значительно улучшить качество при работе с несколькими созависимыми рядами.

Так как VAR модель не приспособлена для работы с рядами, включающими сезонные изменения, предварительно было использовано stl разложение для нивелирования вклада сезонноной компоненты.

Мы попробуем обучить две следующие VAR модели:

1 VAR internal модель будет использовать информацию между временными рядами, соответствующими прибыли различных филиалов, для их непосредственного прогнозирования. Для всех филиалов будет обучена единая модель, учитывающая взаимозависимость сразу всех филиалов.

2 VAR external модель будет для каждого временного ряда, соответствующего прибыли отдельного филиала, использовать внешнюю информацию, включающую кадровые факторы филиалов (число сотрудников и др.), информацию с биржы (курсы акций, доллара и др.) и непосредственно усредненный ряд. Для каждого филиала будет обучена отдельная VAR external модель.

После подбора параметров VAR моделей получим $MAE_{VAR\ internal} = 471$, $MAE_{VAR\ external} = 480$.

Несмотря на использование взаимной информации VAR модели получились не лучше навиного решения и по качеству сравнимы с ARIMA моделями. Как нам кажется, основная причина такого низкого качестве та же, что и у ARIMA моделей, - изменение тренда.

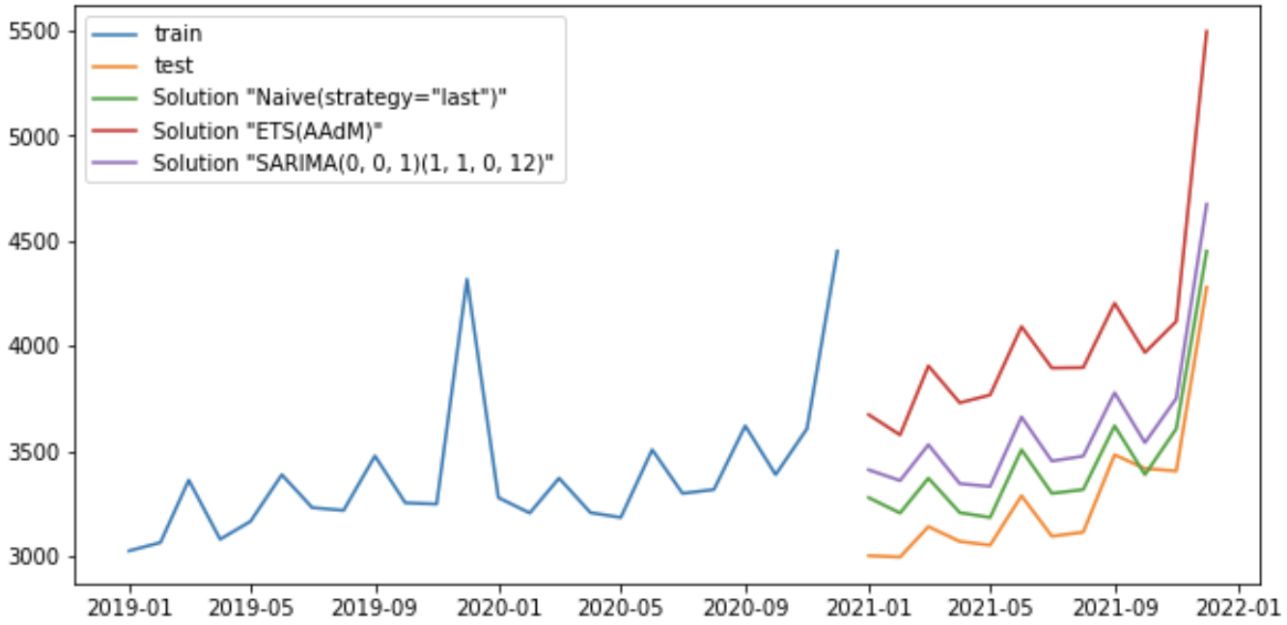


Рис. 10.2: Усредненные ряды по реальным данных и по прогнозам VAR моделей

10.5 Использование классических ML методов

В качестве классических ML методов было принято решение использовать линейную регрессию (Linear Regression) и регрессию основанную на случайном лесе (Random Forest Regressor).

Для каждого типа регрессора для каждого филиала была обучена и применана своя модель. Модели требовалось по значению прибыли, факторов аналогичных факторам для VAR external модели и разностям этих факторов во времени в некотором месяце 2019 года предсказывать прибыль филиала в 2020 году. После, модель применялись аналогичным образом для вычисления целевых значений для 2021 года с помозщью данных за 2020.

В результате были получены следующие значения ошибок $MAE_{LR} = 336$ и $MAE_{RFR} = 315$ для линейной регрессии и регрессии основанной на случайном лесе соответственно.

Рассмотреть усредненный ряд, полученный по предсказаниям ARIMA модели, можно на рисунке 10.3. Видно, что оба прогнозированных ряда по значениям крайне близки к прошлогодним, однако модели дополнительно учитывают некоторую информацию, вносящую свою добавку в предсказания, чем и может объяясняться получение лучшей ошибки по сравнению с наивным подходом.

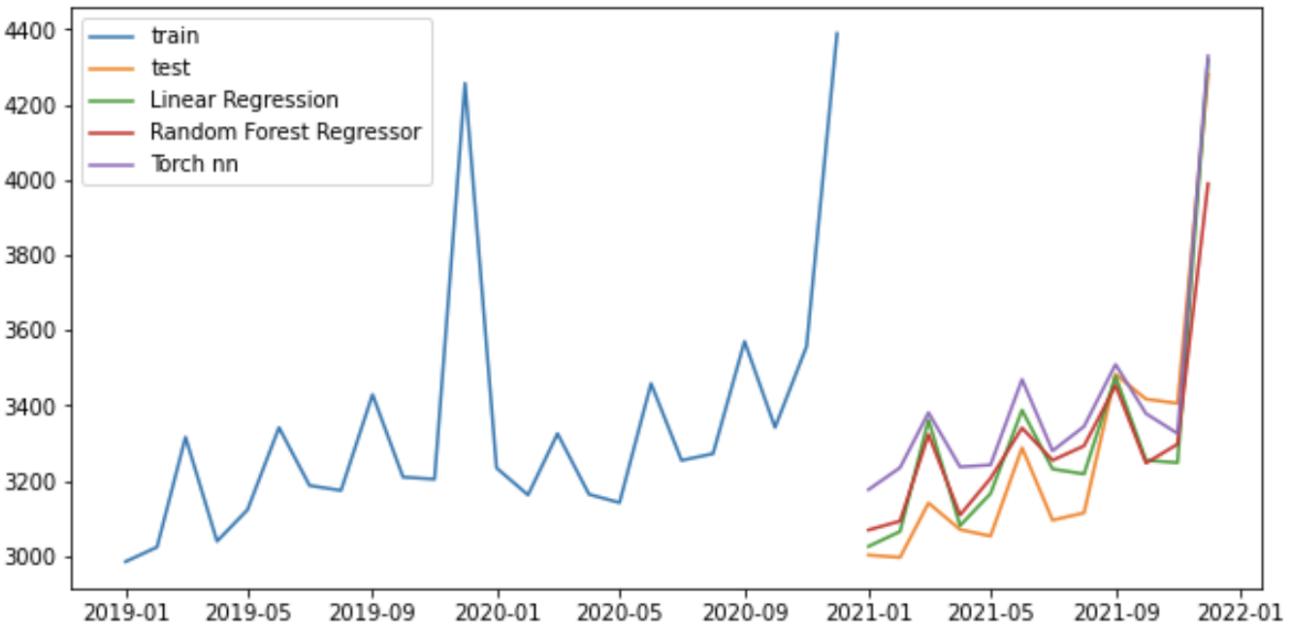


Рис. 10.3: Усредненные ряды по реальным данных и по прогнозам ML моделей

10.6 Использование методов глубинного обучения

В качестве модели глубинного обучения была выбрана модель MLP (Multi Layer Perceptron) с ReLU функциями активации. Более конкретно модель можно представить формулами (1) и (2), где y_0 описывает входные данные, а y^n - результат работы n-слойного MLP. В таких терминах обучение MLP модели - это подбор оптимальных параметров w

$$\mathrm{ReLU}(x) = \max(0, x) \tag{1}$$

$$y_k^{i+1} = \mathrm{ReLU}\left(\sum_{j=1}^{m} w_{ij} y_j^i + b_{ik}\right) \tag{2}$$

11 Заключение (промежуточное)

Хотя наивный подход оказался лучше других по факту исследования данной задачи, это по большей части связано с резким и неожидаемым даже человеку результатом, который могла бы уловить модель на основе машинного обучения, статистические методы выдали более интерпретируемые, близкие интуиции результаты, достойно описывающие процесс.

Список литературы

Dimitros Asteriou and Stephen G. Hall. Arima models and the box–jenkins methodology. *Applied Econometrics (Second ed.). Palgrave MacMillan. pp. 265–286.*, 2011.

Marcus Astley and Olga Cherkashyna. Developing diversity strategies using hr analytics. *Preprint*, 2021. URL https://www.researchgate.net/publication/348442298_Developing_Diversity_Strategies_using_HR_Analytics.

Sandy D Balkin and J Keith Ord. Automatic neural network modeling for univariate time series. *International Journal of Forecasting, 16(4):509–515*, 2000. URL https://www.sciencedirect.com/science/article/abs/pii/S0169207000000728.

R. B. Cleveland, W. S. Cleveland, J. E. McRae, and I. J. Terpenning. Stl: A seasonal-trend decomposition procedure based on loess. *Journal of Official Statistics, 6(1), 3–33*, 2019. URL https://www.wessa.net/download/stl.pdf.

Paul Goodwin. The holt–winters approach to exponential smoothing: 50 years old and going strong. *Foresight: The International Journal of Applied Forecasting*, 2010.

Arindam Banerjee Hardik Goel, Igor Melnyk. R2n2: Residual recurrent neural networks for multivariate time series forecasting. 2017. URL https://arxiv.org/pdf/1709.03159.pdf.

C. E. Holt. Forecasting seasonals and trends by exponentially weighted averages. *O.N.R. Memorandum No. 52*, 1957.

B. Letham and Taylor S.J. Prophet: forecasting at scale. *facebook research*, 2017. URL https://research.fb.com/prophet-forecasting-at-scale.

Edouard Ribes, Karim Touahr, and Benoit Perthame. Employee turnover prediction and retention policies design: a case study. *CoRR, Volume 10*, 2017. URL https://arxiv.org/pdf/1707.01377.pdf.

Lionel P. Robert, Casey Pierce1, Liz Morris, Sangmi Kim, and Rasha Alahmad. Designing fair ai for managing employees in organizations: A review, critique, and design agenda. *Human–Computer Interaction, Volume 35*, 2020. URL https://arxiv.org/pdf/2002.09054.pdf.

Alex Sherstinsky. Fundamentals of recurrent neural network (rnn) and long short-term memory (lstm) network. *Elsevier "Physica D: Nonlinear Phenomena"journal, Volume 404, March 2020: Special Issue on Machine Learning and Dynamical Systems*, 2018. URL https://arxiv.org/pdf/1808.03314.pdf.