**I. Описание работы**

За последний год слово «биткоин» – стало синонимом цифровой экономики. Не утихают споры является ли он прообразом валюты будущего, несёт ли в себе какую-то ценность (value), способен ли защитить от инфляции, является ли чисто спекулятивным инструментом и не является ли просто очередным «пузырём» (<https://www.cnbc.com/2018/01/16/bitcoin-is-a-pyramid-scheme-warns-ex-wells-fargo-ceo-dick-kovacevich.html>). С юридической точки зрения до сих не закрыт вопрос следует ли принимать его, как платёжное средство, или запретить для конвертации в товары внутри страны. Стоят вопросы по регуляции и налогообложению действий майнеров, которые своими действиями по сути создают деньги из воздуха (<https://www.cnbc.com/2018/02/21/everything-you-need-to-know-about-bitcoin-and-your-taxes.html>). Всё вышеперечисленное позволяет утверждать, что по сравнению с ситуацией ещё пару лет назад, зарождением биткойна в 2009-ом году, внимание к криптовалюте заметно увеличилось. Прежде скептические реакции о возможном будущем биткойна сменились на более серьёзную аналитику. Многие сайты начали выпускать ежедневные обзоры по крипторынкам. Даже в таких серьёзных изданиях, как Bloomberg, в день выходит несколько десятков новостей о криптовалюте. Появились специализированные сайты, которые пишут только о криптовалюте. Всё больше людей начинают следить за вводом новых протоколов, начинают разбираться в принципе работы блокчейна, в том, что такое lighting, зачем нужен segwit, как грядущий fork повлияет на цену криптоактива. Но как все эти знания могут помочь опытному инвестору? В данной работе я попытаюсь доказать или опровергнуть факт, что новостной фон вокруг биткойна содержит возможность для создания роботизированного алгоритма для предсказания движения криптовалюты в следующий момент.

В работе, используя базовые методы NLP (Natural Language Processing), я оцениваю, какой вклад дают публикации на новостных сайтах в будущую цену биткойна. В качестве новостных сайтов я выбрал пять наиболее известных сайтов: Bitcoin.com, Coindesk.com, Cointelegraph.com, Bloomberg, Forbes. Биткойн выбран, как одна из основных криптовалют по капитализации. Моей задачей являлось предсказать рост или падение цены биткойна после публикации новости. При этом результаты моей работы можно распространить дальше на другие криптовалюты или другие новостные сайты. Также результаты работы можно использовать для создания собственных торговых роботов или просто для торговли на бирже. Кроме этого результаты работы являются проверкой теории эффективности рынка утверждающей, что вся публичная информация уже учтена в цене криптовалюты. В случае если вся информация уже учтена в цене акции, построение подобных торговых роботов невозможно. В работе будут проанализированы, как дневные, так и минутные данные. Анализ минутных данных позволит говорить о внутридневных корреляциях и о том, как быстро рынок ребалансируется при публикации новостей.

Новаторство данной работы заключается в том, что ранее не было исследовано влияние новостного потока на курс криптовалют. Методы, применяемые мной для анализа, позволяют понять, на сколько быстро новости учитываются в цене биткойна. Анализ несколько разных сайтов позволяет понять, какой источник формирует мнение для трейдеров и является более авторитетным.

**II. Биткойн, как криптовалюта**

В данном параграфе я рассмотрю основные аспекты работы биткойна и некоторые общие свойства, присущие другим криптовалютам и блокчейну. Я дам описание, как работает биткойн, как происходят транзакции, как работает крипто-кошелёк, что такое алгоритм консенсуса. Также я попытаюсь дать некоторое представление о том, почему биткойн является валютой, раскрою некоторые из взглядов на то является ли он пузырём или нет. Биткойн является инновацией не только благодаря тому, что является цифровой валютой, но и из-за блокчейна, который лежит в его основе. Поэтому при описании биткойна я так же обращаю внимание на некоторые аспекты работы блокчейна и полезных свойств, которые заложены в нём. Следует отметить, что многие свойства биткойна могут быть обобщены в целом на любую криптовалюту. В силу этого моя работа по описанию биткойна имеет практический интерес для тех, кто хочет разобраться, как работает этот финансовый инструмент.

**Биткоин**

Биткоин – это цифровая платёжная система, состоящая из цепи транзакций. В основе этой цепи лежит алгоритм консенсуса между участниками системы. Под консенсусом в сети мы имеем в виду, что транзакция в сети считается валидной, если за неё проголосовала большая часть пользователей системы. Биткоин является первой peer-to-peer платёжной системой, полностью поддерживаемой пользователями, без участия регулирующих органов и посредников. («Pantera Primer», Ronald A.Glantz, 2014) Что такое peer-to-peer сеть? Также её называют p2p, одноранговой, пиринговой или децентрализованной. P2P – это компьютерная сеть, созданная поверх основной сети (в нашем случае интернета) и основанная на равноправии участников. В такой сети отсутствуют выделенные серверы, а каждый узел (peer) является, как клиентом, так и выполняет функции сервера). (<https://en.wikipedia.org/wiki/Peer-to-peer>)

С точки зрения пользователя всё намного проще, биткойн просто является валютой в интернете. Хотя стоит отменить, что кроме того биткойн является самой перспективной и надёжной книгой учёта транзакций в мире. Надёжность и сохранность достигается за счёт того, что все данные много раз скопированы и хранятся на компьютерах пользователях сети. На момент написания этой статьи в интернете присутствовало 1592 криптовалюты (<https://coinmarketcap.com/all/views/all/>), при этом биткойн по капитализации занимал лидирующее место с объёмом в 37% от всех существующих криптовалют (<https://coinmarketcap.com/charts/>).

Опишем принцип, по которому работает биткойн. Для работы с биткойном нужен специальный компьютерный софт или мобильное приложение, которые содержат информацию о цифровом кошельке пользователя и позволяют получать и отправлять биткойны через них. Получение и отправление биткойнов происходит посредством добавления новых записей в специальный электронный журнал транзакций, который называется «блокчейн». Этот журнал транзакций находится в публичном доступе, и любой пользователь может при желании скачать себе информацию обо всех транзакциях всех пользователей интернета. Подлинность каждой транзакции обеспечивается электронной подписью отправителя. Таким образом только пользователь кошелька управляет тем, кому и сколько биткойнов он отправит. Также любой пользователь может осуществить процессинг транзакций в блокчейне, используя мощности своего компьютера или специальных серверов. Этот процессинг называют «майнингом». В вознаграждение за майнинг система даёт некоторое количество биткойнов.

Итак, каждый пользователь кошелька имеет публичный и приватный ключи. Аналогом публичного ключа является банковский счёт. Вы можете переводить деньги на любой публичной ключ. Но только обладатель приватного ключа может перевести/потратить биткойны. Все публичные ключи лежат в открытом доступе, но мы не знаем, кому принадлежит какой ключ. Поэтому говорят, что адреса биткойнов псевдоанонимны. При создании кошелька и перевода на него биткойнов, кошелёк сигнализирует в блокчейн, что на него поступили деньги. Кошелёк генерирует два ключа – публичный и приватный (приватный может не генерироваться, если кошелёк создан на бирже, например, Bitsamp). Если вы захотите заплатить за какую-то услугу, то вы идёте к продавцу и говорите, что с вашего публичного ключа ему прийдут биткойны. Дальше мы идём в кошелёк или на биржу и переводим средства с нашего публичного ключа на публичный ключ продавца. Эта транзакция попадает в p2p сеть, которая проверяет её, смотрит, есть ли у вас деньги, и отправляет для записи в узлы сети. Таким образом, ваша транзакция попадает майнеру, который добавляет её в «блок». Блок состоит из серии транзакций разных пользователей. Блок передаётся в сеть дальше, где он попадает в общий журнал транзакций, т.е. в блокчейн. Теперь ваши деньги переведены продавцу. («Pantera Primer», Ronald A.Glantz, 2014)

**Преимущества биткойна**

* Свобода в совершении транзакций. Вы можете посылать деньги любому участнику сети в любом объёме в любое время. Сеть работает без выходных, без географических границ, без других финансовых ограничений.
* Маленькая комиссия за транзакцию. Теоретически процессинг транзакций может быть для вас бесплатным. На практике для ускорения процессинга вы платите небольшую сумму майнерам.
* Привлекательность для микротранзакций. Так как комиссия за транзакцию очень маленькая, то биткойн может использовать в развивающихся странах продавцами для осуществления микроплатежей.
* Прозрачность сети. Вся информация о платежах находится в блокчейне. В любой момент кто угодно может проверить и верифицировать ваш платёж.
* Нейтралитет. Так как биткойн представляет из себя p2p сеть, т.е. децентрализирован, то никто не может манипулировать и контролировать его протокол.

(«Pantera Primer», Ronald A.Glantz, 2014)

**Биткойн, как денежная единица**

Ценность биткойна определяется возможностью использовать его в виде денежной единицы. Его принимают в виде платёжного средства, на него можно купить что-то, его можно обменять на другие денежные единицы. Кроме того, биткойн обладает и другими свойствами денег: продолжительный срок службы, портативность, возможность обменять на что-то, редкость (точнее заложенная в нём ограниченность максимального количества возможных добытых единиц), делимость на более мелкие единицы, несёт в себе некоторую цену (store of value), узнаваемость. Отличие от других денежных единиц состоит в том, что биткойн не несёт в себе никакие физические свойства (как золото или серебро) или его ценность не базируется на веру в органы власти (как у обычных фиатных денег). Биткойн целиком основан на математических свойствах. Так как биткойн обладает всеми свойствами денег, то его цена в главную очередь определяется принятием его другими контрагентами. Со временем всё больше контрагентов будут принимать биткойн в качестве оплаты за услуги и товары, и его ценность от этого будет только расти.

Как и у другой денежной единицы цена биткойна определяется спросом и предложением. Предложение биткойна ограничено математически – существует ограничение на общее количество биткойнов, которое будет сгенерировано системой. Спрос на биткойны определяется пользователями и участниками сети. В последнее время цена биткойна часто имела сильную волатильность из-за резкого наплыва и последующего исчезновения новых участников рынка, которые двигали цену вверх или вниз. В будущем, когда капитализация биткойна достаточно вырастет он будет менее подвержен подобным шокам.

(«Pantera Primer», Ronald A.Glantz, 2014)

**Является ли биткойн пузырём**

To be done

**Легализация биткойна**

To be done

**Алгоритм Консенсуса**

<https://habrahabr.ru/company/bitfury/blog/327468/>

**DAO**

<https://medium.com/universablockchain/decentralized-autonomous-organization-what-is-a-dao-company-eb99e472f23e>

**III. Источники данных**

Для моделирования эффекта выхода публикации на курс биткойна я буду использовать несколько источников новостей. Всего будет рассмотрено пять разных сайтов: Bitcoin.com, Coindesk.com, Cointelegraph.com, Bloomberg, Forbes. При отборе источников я руководствовался в первую очередь узнаваемостью и цитируемостью источника. Так первые три сайта появляются первыми в выдаче в Яндексе и Google по запросам «биткойн» и «биткойн новости». Bloomberg выбран, как один из основных источников новостей о финансовых рынках. Forbes я взял, как альтернативу Bloomberg. По содержанию статьи на Forbes являются более полными и предоставляют более широкую аналитику, чем публикации на сайтах bitcoin или coindesk, но по характеру контента Forbes больше похож на них, потому что при публикации контента придерживается модели «contribution model» (статьи написаны самими читателями или независимыми журналистами).

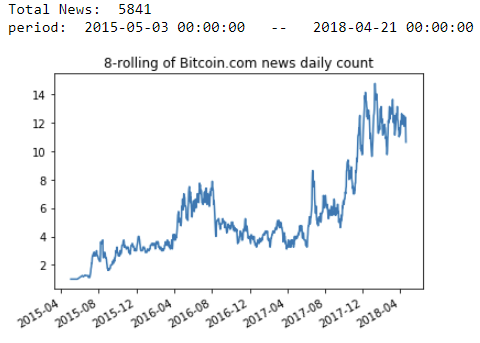
В качестве предварительной гипотезы я делаю предположение, что контент на сайтах Bitcoin.com, Coindesk.com, Cointelegraph не будет оказывать значимого влияния на курс биткойна, т.е. по содержанию контента мы не сумеем предсказать движения криптовалюты. Предположение сделано на основе того, что данные сайты не предоставляют аналитику в реальном времени и по больше части публикуют новости post factum. Таким образом, новости будут сильно коррелировать с динамикой внутри дня, но не оказывать влияния на движение валюты на следующий день. Аналогичную ситуацию мы будем наблюдать с сайтом Forbes.com; отличие состоит в том, что высокая посещаемость сайта может создавать критическую массу и всё-таки оказывать влияние на движение биткойна. Bloomberg предоставляют аналитику финансовых рынков в реальном времени и является основным источником информации для большинства трейдеров, поэтому для него я ожидаю увидеть значимую корреляцию с движением криптовалюты. Кроме того, данные о новостях с сайта Bloomberg доступны с точностью до минуты публикации, что даёт возможность исследовать вклад публикации в цену биткойна в следующую за публикацией минуту.

Ниже даны подробные описания каждого источника данных с краткими пояснениями по структуре исходных данных.

**Bitcoin.com**

Компания, предоставляющая сервисы по работе с Bitcoin Core и Bitcoin Cash. Через сайт можно покупать и продавать криптовалюты, а также работать с криптокошельком. На самом сайте содержатся подразделы, где можно посмотреть новости из крипто рынка, построить графики для анализа цены крипто валюты, поиграть в онлайн казино на крипто валюту. Также через сайт существует возможность поучаствовать в майнинге криптовалюты посредством участия в облачной бирже мощностей (<https://en.wikipedia.org/wiki/Bitcoin.com>)

Главный офис компании находится в Токио. CEO Роджер Вер. Домен функционирует с 2014-ого года. Сайт занимается популяризацией биткойна. На апрель 2018-ого года сайт занимал первое место в Яндексе и первое место в Google по запросу «bitcoin news».



**Coindesk.com**

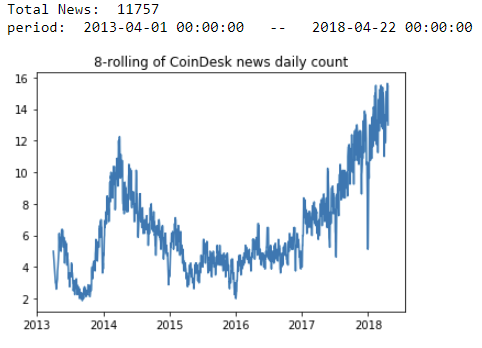
Сайт был запущен в мае 2013-ого года. Портал специализируется на новостях о криптовалютах. Сайт основал Шакил Кан (Shakil Khan), в последствии сайт приобрела компания Digital Currency Group. (<https://en.wikipedia.org/wiki/CoinDesk>). Посещаемость сайта в день составляет 10 миллионов уникальных человек, которые в целом делают 50 миллионов посещений. Coindesk присутствует в социальных сетях: твиттере, фейсбуке, линкедине, - где на него подписаны более 600 000 подписчиков. (<https://www.coindesk.com/about/>)

С сентября 2013-ого года сайт публикует Bitcoin Price Index (CoinDesk BPI). Индекс является средней ценой биткойна нескольких крипто бирж. Индекс часто цитировался в ведущих новостных изданиях, таких как The Wall Street Journal, Financial Times, Bloomberg, CNBC. (<https://www.coindesk.com/about/>)

Кроме того, CoinDesk проводит ежегодную конференцию Consensus summit. Это самая большая встреча в мире, посвящённая технологии блокчейн. Конференция проходит в городе Нью-Йорк. В мае 2017-ого года на саммит было продано свыше 2700 билетов. CoinDesk также проводит саммит Consensus: Invest для институциональных инвесторов, хедж фондов, банков и интересующихся инвесторов, где обсуждается, как инвестировать, где хранить, как оценивать стоимость крипто активов.

Другой известной частью деятельности CoinDesk является отчёт “The State of Blockchain” – самый подробный анализ рынка и перспектив технологии блокчейн

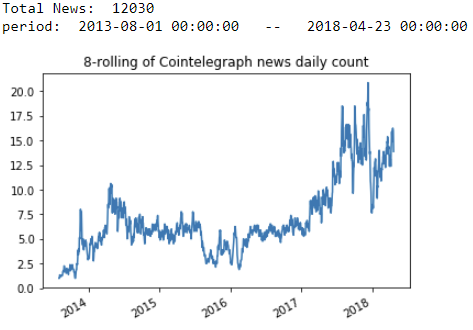
На апрель 2018-ого года сайт входил в топ 10 по запросу «bitcoin news» в Яндексе и занимал второе место в Google по тому же запросу.



**Cointelegraph.com**

Сайт существует с 2013-ого года. Портал является независимый источником новостей о криптовалюте, блокчейне и децентрализованных приложениях. На сайте есть новости, анализ, экспертные колонки и информация, касающаяся Финтеха, Блокчейна и Биткойна. (<https://www.crunchbase.com/organization/cointelegraph>)

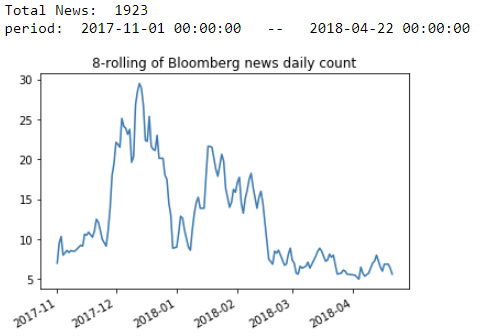
На апрель 2018-ого года сайт входил в топ 10 в Google и в топ 5 в Яндексе по запросу «bitcoin news».



**Bloomberg.com**

Компания была основана в 1981-м году Михаэлом Блумбергом (Michael Bloomberg). Компания должна была стать технологическим поставщиком финансовой аналитики и информации. Сейчас Bloomberg – один из главных глобальных поставщиков круглосуточной информации по финансовым данным и другой финансовой информации, включая текущие и исторические цены на рыночные активы, финансовые отчётности, финансовые новости, аналитику, а также общие новости и новости спорта. Блумберг представляет новости на своей собственной разработанной платформе – терминале – а также на своём ТВ канале (Bloomberg Television), на радио (WBBR), в журналах (Bloomberg Businessweek, Bloomberg Markets, Bloomberg Pursuits) и на сайте (Bloomberg.com). (<https://www.investopedia.com/terms/b/bloomberg.asp>, <https://en.wikipedia.org/wiki/Bloomberg_L.P>.)

В целях анализа использовались только новости с Блумберг, которые можно найти на сайте по запросу «биткойн» (<https://www.bloomberg.com/search?query=bitcoin>). К сожалению, Блумберг не даёт возможности вытаскивать все существующие старые новости в силу технических ограничений портала. Поиск по старым новостям устроен таким образом, что индексируются единичные самые популярные новости и нет возможности выгрузить все существовавшие публикации. В силу этого ограничения для анализа были доступны только новости, начиная с ноября 2017-ого года. Но для каждой новости была известна точная минута публикации, поэтому эти данные можно использовать для анализа связи с минутными данными по цене биткойна.

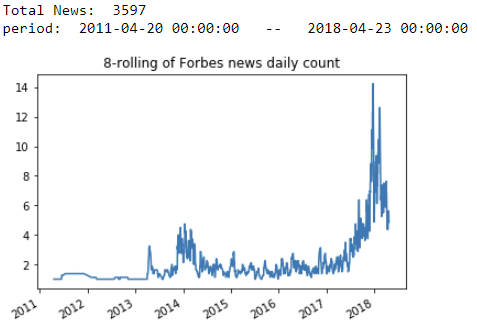


**Forbes.com**

Forbes.com является частью Forbes digital, который в свою очередь является частью Forbes Media LLC. Сайт был основан Дэвидом Чарбаком в 1996-ом году. Так же как и журнал сайт занимается публикацией новостей и статей о долларовых биллионерах, их жизненным увлечениям и об их движимом и недвижимом имуществе. Но на сайте также присутствует много новостей из мира бизнеса и политики. (<https://en.wikipedia.org/wiki/Forbes>)

Forbes.com использует «модель контрибуции» при отборе статей на сайт. То есть контент на сайт создают и публикуют «контрибьюторы». Всем контрибьюторам сайт платит за публикации на основе посещаемости их статей. Всего новости на сайте публикует свыше 2500 человек при этом некоторые из авторов заработали за свою деятельность свыше $100 000 (<https://www.journalism.co.uk/news/the-forbes-contributor-model-technology-feedback-and-incentives/s2/a554255/>).

В целях данной работы использовались новости, которые можно найти на сайте Forbes.com при использовании поискового запроса «bitcoin». Полный список новостей можно посмотреть по ссылке <https://www.forbes.com/search/?q=bitcoin>



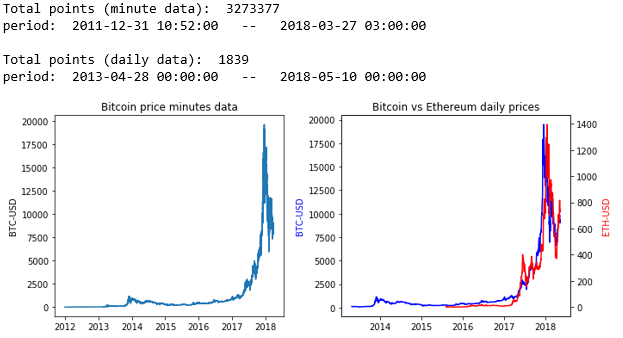
**Курсы криптовалют**

Дневные курсы биткойна собраны с сайта coinmarketcap. Coinmarketcap является главным поставщиком данных о ценах на криптовалюты. И по данным сайта Alexa он является 44м самым посещаемым сайтом в США

(<https://news.bitcoin.com/8-alternatives-to-coinmarketcap/>) (<https://coinmarketcap.com/currencies/bitcoin/historical-data/?start=20130428&end=20180511>)

Минутные данные с курсом биткойна собраны пользователями платформы Kaggle в исследовательских целях. Данные доступны по паре BTC-USD по двум биржам bitstamp и coinbase

<https://www.kaggle.com/mczielinski/bitcoin-historical-data/data>



**IV. Сбор и подготовка данных**

Сбор, обработка и последующий постпроцессинг данных сделаны на Python и доступны в приложении к диплому в виде Jupyter ноутбуков.

**Сбор новостей**

Сбор новостей проходил с помощью скриптов, написанных на языке Python. При обращении к сайту использовалась библиотека requests. Парсинг новостей из html и xml формата проходил с помощью библиотеки lxml и её модулей html и etree. При парсинге сначала читалась страница из поиска новостного сайта со списком релевантных новостей. Затем происходил переход по ссылке с каждой новостью, чтобы прочитать её текст. При чтении списка новостей с сайта Forbes использовался внутренний API сайта. В остальном для каждого сайта писался индивидуальный загрузчик в силу того, что структура новостей на каждом сайте своя.

После чтения html кода с новостью из него удалялась вся нерелевантная информация. Происходила очистка от html тэгов, удалялись все непечатные символы, реклама, картинки. Помимо текста для некоторых сайтов была выгружена дополнительная информация, такая как заголовок новости, имя автора, краткое описание новости, тип новости (новость, блог, аудио, колонка, фото-эссе…), количество лайков и шэров (shares).

**Препроцессинг новостей**

Препроцессинг новостей включал в себя стандартную обработку, характерную для задач Natural Language Processing. Похожий процесс проделывался во многих работах. Например? в Sentiment “Analysis of Twitter Data for Predicting Stock Market Movements” (<https://arxiv.org/pdf/1610.09225.pdf>)

1. Разделение новости на заголовок и текст новости. В дальнейшем из заголовка и из самой новости строились отдельные переменные.
2. Объединение новостей, вышедших в один день. Новости, которые вышли в один день, объединялись в один сплошной текст. В данной работе я не ставлю цель определить наиболее повлиявшую на курс биткойна новость. Предполагается, что на курс биткойна повлиял весь выходивший контент. В дальнейшем в тексте при упоминании слова «новость» я буду считать, что это как раз и есть текст, объединённый из всех новостей, вышедших в один день.
3. Токенизация. Процесс разбивки слов на отдельные токены. Каждое слово представляет единую сущность. Новости, вышедшие в один день, представляют из себя вектор этих слов, токенов.
4. Удаление стопслов. При проведении анализа из новостей выкидывались стоп слова, характерные для английского языка. Также выкидывались знаки пунктуации, html тэги и другие спецсимволы.
5. Замена слов с помощью регулярных выражений. С помощью регулярных выражений заменялись некоторые слова и специальные обозначения. Так все встречающиеся аббревиатуры “R&D” были заменены на “RnD”, цифры с припиской “m” и “bn” были заменены на “millions” и “billions”. Знаки “$” и “€” были заменены на “USD” и “EUR”.
6. Приведение слов к их основам. Отдельные токены, слова были заменены на их основы. То есть из слов были удалены окончания, соответствующие мужскому, женскому роду, множественному числу, прошедшему времени и тп. При лематизации слов использовались функции из пакета nltk.
7. Превращение токенов в N-граммы. В некоторых случаях для улучшения качества модели были использованы n-граммы, где n равнялось 2 или 3. N-граммы представляют из себя множество текстовых элементов, где учитываются не только индивидуальные слова (токены), но и словосочетания из n-слов, идущих друг за другом. Так 2-граммы из фразы “New currency, same blockchain?” представляют из себя следующий массив: “New”, “currency”, “same”, “blockchain”, “new currency”, “currency same”, “same blockchain”. В дальнейшем под словом токен я подразумеваю не только отдельные слова, но и n-граммы из той же новости.
8. Отбрасывание редких n-граммов. Если словосочетания встречались в корпусе новостей реже 20 раз, то я отбрасывал их, чтобы они не попадали в дальнейший анализ.

**TF-IDF Преобразование**

После преобразования новостей в n-граммы для дальнейшего использования этих данных в регрессиях и классификаторах было осуществлено TF-IDF преобразование. Данное преобразование позволяет оценить важность отдельных n-грамм документа в коллекции документов. Преобразование учитывает не только, как часто слово встречается в документе, но и как часто слово встречается во всей коллекции новостей. Делается это для того, чтобы слова, которые встречаются часто во многих документах, в итоге имели меньше вес, чем более редкие слова.

Здесь TF отвечает за частоту встречаемости конкретного токена в корпусе слов:

, где n() – количество слов. – j-ая новость, – i-ое слово. Таким образом, мы делим количество раз, которое токен встречается в новости на общее количество слов в новости .

Часть IDF отвечает за то, как часто слово встречается во всём корпусе новостей. Мы делим количество документов в корпусе |D| на количество документов, где встречается слово , которое равняется . Дальше от этого отношения мы берём логарифм.

Разложение TF-IDF неоднократно использовались при анализе и моделировании биржевых курсов. Например, можно посмотреть на результаты работы Predicting Stock Price Movement Using Social Media Analysis, написанную Derek Tsui (<http://cs229.stanford.edu/proj2016/report/Tsui-PredictingStockPriceMovementUsingSocialMediaAnalysis-report.pdf>)

**Препроцессинг цены биткойна**

Для анализа цены биткойна использовались два метода. Первое – это стандартное представление доходности в виде разности логарифмов:

Второе представление использовалось для классификации движения курсы биткойна. Задачей данной работы является предсказание периодов роста и падения биткойна. Для этого мы сравниваем доходность биткойна. Если она больше нуля, то биткойн растёт, если меньше нуля, то падает.

Здесь я следую разложению из работы (<http://www.public.asu.edu/~hdavulcu/WI15.pdf>)

**V. Моделирование**

Моделирование проводилось на языке Python в Jupyter notebook с использованием библиотеки scikit. Код доступен в приложении к диплому.

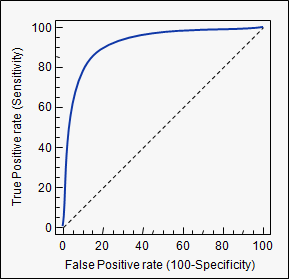
**Проверка качества модели**

При моделировании проверялось качество предсказания переменной . Т.е. решалась задача классификации. В качестве метрики корректности предсказаний использовалась ROC-AUC. (<https://en.wikipedia.org/wiki/Receiver_operating_characteristic>). Кривая ROC строится путём сравнения true positive rate (TPR) с false positive rate (FPR) при различных вариантах отсечения по скорам. Ниже описан способ построения ROC кривой.

Для начала приведём определения TPR и FPR:

Для построения ROC кривой мы указываем для каждого наблюдения из выборки некоторый скор . Этот скор соответствует мнению модели о том, что данное наблюдение . Все наблюдения мы ранжируем по скорам, получая последовательность:

Теперь последовательно для каждого скора из последовательности мы считаем метрики TPR и FPR. Таким образом, для каждого наблюдения мы получаем две точки и . Эти точки являются координатами для графика ROC:



(источник картинки <https://stats.stackexchange.com/questions/132777/what-does-auc-stand-for-and-what-is-it>)

ROC-AUC (area under the curve) – это площадь под кривой ROC. Посчитанная метрика AUC может быть интерпретирована, как вероятность того, что классификатор покажет скор случайного выбранного наблюдения «движения вверх» выше, чем скор случайного наблюдения «движение вниз».

**Модель Ridge**

В качестве одной из основных моделей для классификации использовалась регрессия Риджа. Данная регрессия удобна тем, что является линейной, то есть её проще оценить и интерпретировать. Она позволяет использовать сразу много переменных и при этом сохраняет высокую скорость решений при увеличении размерности задачи. От стандартной линейной регрессии она отличается тем, что добавляет пенальти для больших коэффициентов, что препятствует переобучению модели. Общий вид регрессии Риджа следующий:

**Модель ExtraTrees**

Другой метод, который я буду использовать при моделировании. Он предоставляет альтернативу линейной регрессии при сохранении скорости работы, необходимой при обработке большого количества текстовой информации. Метод был предложен в 2005-м году и является дополнением к методу random forest. Как и random forest, он представляет собой ансамбль моделей деревьев решений, каждое из которых построено с помощью техники бутстрап (bootstrap). При этом на каждом шаге при делении дерева на ветви используется случайный набор признаков (особенность random forest) и случайная точка для деления каждого признака (особенность extra trees).

Подробнее о методе можно почитать в оригинальном документе Extremely randomized trees (<https://orbi.uliege.be/bitstream/2268/9357/1/geurts-mlj-advance.pdf>)

**Обучение и проверка**

Весь исследуемый временной интервал делился на обучающую выборку и на валидационную выборку. Обучающей выборкой являлись первые 95% наблюдений. Валидационной оставшиеся 5%. Данный подход деления на тестовую и валидационную выборку является стандартным и используется, к примеру, в работе <http://www.public.asu.edu/~hdavulcu/WI15.pdf>

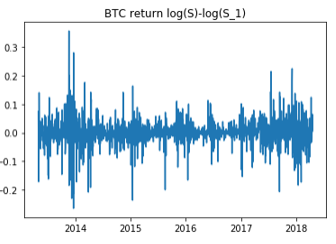
Подбор параметров моделей проводился с помощью кроссвалидации на 10ти фолдах. То есть весь исследуемый временной отрезок делился на 10 частей. Далее последовательно обучение происходило на 9-ти частях этого отрезка и проверка качества на 10-ом. Полученные результаты работы модели для каждого из 10-ти фолдов усреднялись.

**VI. Результаты**

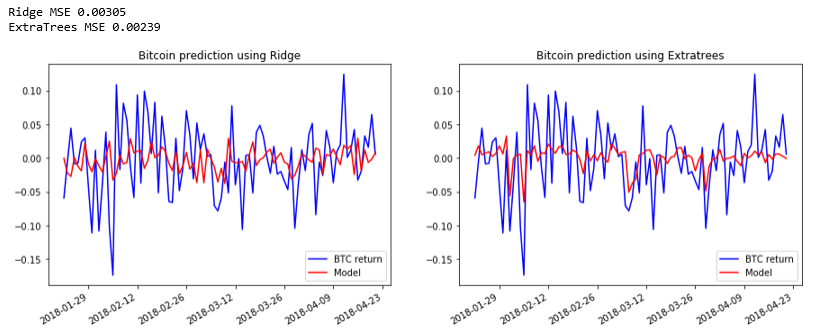
В первой части данного раздела я бы хотел показать результаты, которые можно получить, если моделировать курс биткойна, а не его направление движения. Забегая вперёд, только данных из новостей недостаточно, чтобы предсказывать скачки в цене биткойна. Поэтому в последующих частях моделироваться будет направление движения, то есть переменная .

**Моделирование доходности биткойна**

Для моделирования доходности биткойна применим к цене биткойна преобразование в виде разности логарифмов текущей цены и цены в предыдущий день, как это было описано в предыдущем разделе. Полученный граффик будет напоминать белый шум. При этом ясно наблюдаются периоды сильной волатильность биткойна, особенно в 2014-ом году и в 2017-2018-ом годах.



Следующим этапом возьмём новости с сайта Coindesk и попробуем построить регрессии, используя методы Риджа и Extra Trees. При моделировании будем смотреть на новости, которые выходили в тот же день, когда цена биткойна изменилась. Тем самым мы попробуем установить часто ли в новостях пытаются объяснить изменения цены биткойна какими-то факторами. Ниже на графике я привожу сравнение доходности курса биткойна и результатов моделей. Мы видим, что в целом модель ловит некоторые тренды, но качество моделей тем не менее довольно плохое. Модель ExtraTrees сумела схватить несколько больших провалов, но не один пик так и не был объяснён. Таким образом, мы можем сделать вывод, что новости не всегда выходят после скачков цены биткойна и не всегда содержат объяснения этим скачкам.



В регрессии Риджа мы можем посмотреть на самые большие коэффициенты в модели, чтобы определить наиболее значимые слова. Как и предполагалось наиболее значимые новости содержат слова “drop” и “price drop”. Что скорее всего говорит о том, что чаще всего новости выходят после падения цены биткойна, где Coindesk пытается объяснить причину этого падения. Среди слов, дающих положительный вклад, мы можем выделить “high”, “miner” и “patent”. Можно предположить, что цена биткойна растёт, когда мы слышим положительные новости со стороны развития технологии биткойна (появления новых патентов) или позитивные новости о майнерах, которые увеличили свои мощности, тем самым увеличив сеть блокчейна.

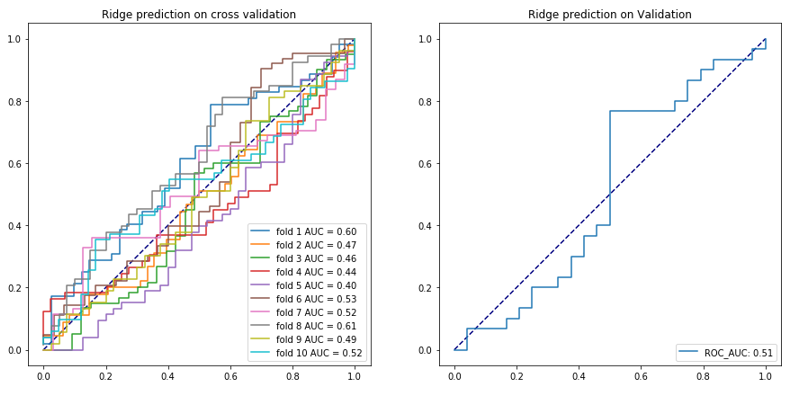
|  |  |
| --- | --- |
| drop | -0.0547605243076 |
| price drop | -0.0445815059531 |
| drop usd | -0.0433770776923 |
| fall | -0.0407284926075 |
| high | 0.0390665698758 |
| miner | 0.0343074445839 |
| official | -0.0325386999389 |
| pass | 0.0308081809046 |
| patent | 0.0307477307281 |

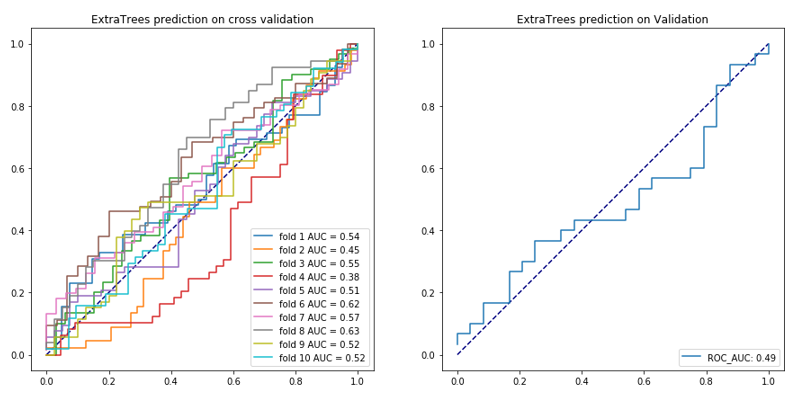
В целом данная модель показывает, что очень сложно предсказать цену биткойна. Новости не успевают объяснять все пики и провалы криптовалюты. Для дальнейших исследований эту модель можно было бы улучшить, для чего я предлагаю исследовать влияние лагов в переменной новостях (возможно, некоторые новости выходят через несколько дней после падения или роста биткойна). Также можно было бы попробовать сделать совместные модели с нескольких новостных сайтов. Другим пространством для улучшения модели могло бы быть уменьшение горизонта предсказания до недели или нескольких дней. Вполне возможно, что новостной фон меняется очень быстро, поэтому сложно использовать закономерности найденные в старых новостях для предсказания цены биткойна на несколько месяцев вперёд.

**Моделирование Bitcoin**

Теперь вернёмся к моделированию направления движения цены биткойна. В качестве целевой переменной у меня будет выступать . Будем смотреть, как новость предсказывает целевую переменную. При валидации будем использовать TimeSeriesValidation из библиотеки scikit-learn.

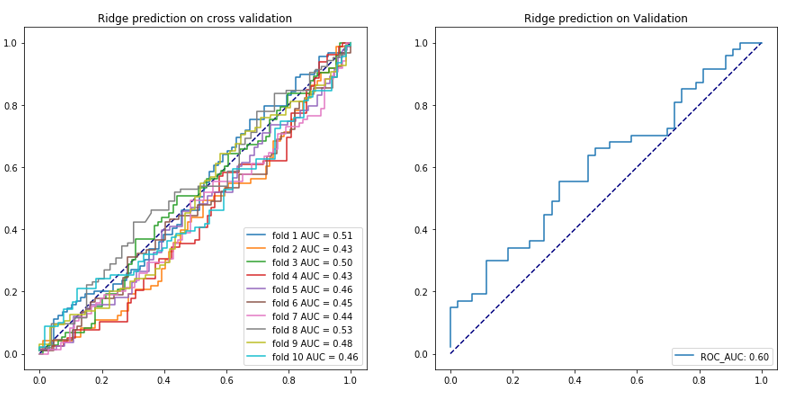
Несмотря на низкое значение AUC на валидационной выборке, мы можем видеть значения AUC выше 50% для фолдов с 5-ого по 10-ый для метода ExtraTrees. И хотя значения ROC-AUC на этих фолдах не очень большое, мы всё равно можем сделать предположение о наличии некоторой предсказательной способности у данного датасета.

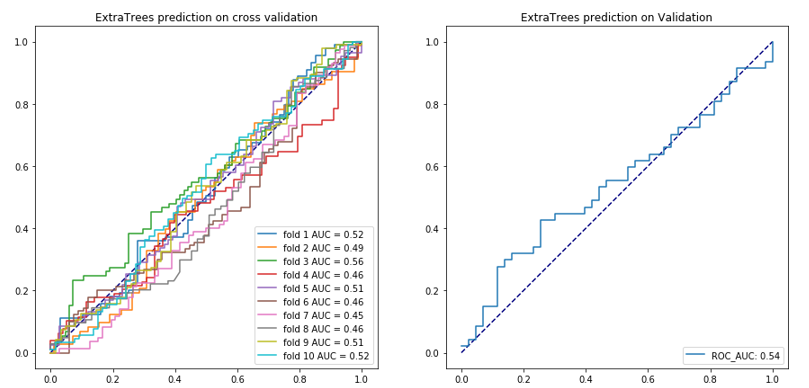




**Моделирование Coindesk**

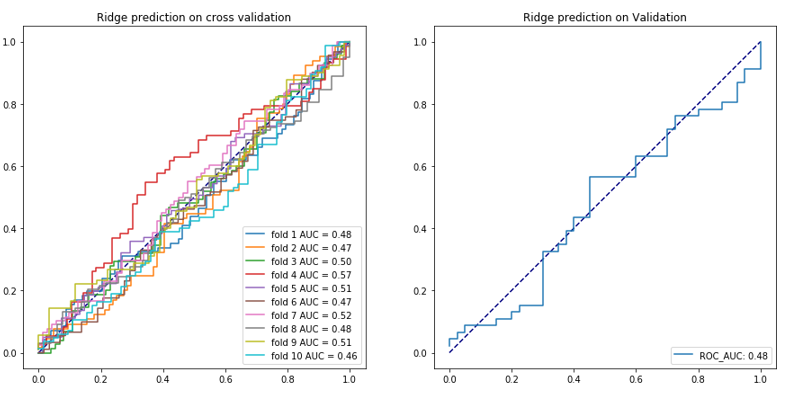
Практически на всех фолдах кроме валидационного датасет показал плохую предсказательную способность. Возможно, предсказательная способность сайта на валидации улучшилась по сравнению с историческими периодами, т.е. новости начали нести более существенный вклад в цену криптоактива, но т.к. зависимость не подтверждается на исторических данных, то мы не могли бы использовать её для торговли.

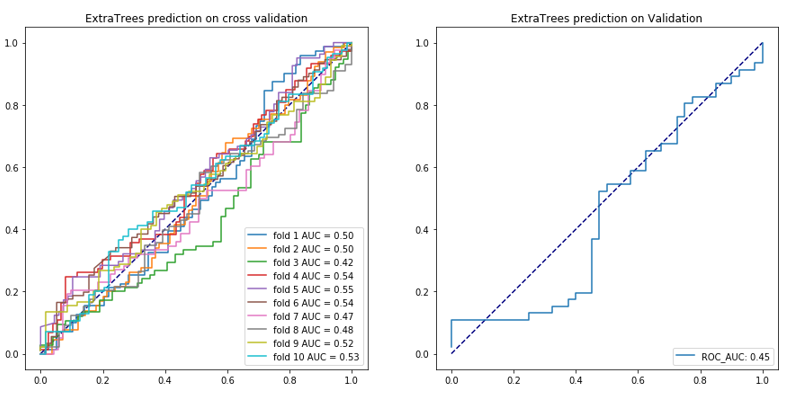




**Моделирование CoinTelegraph**

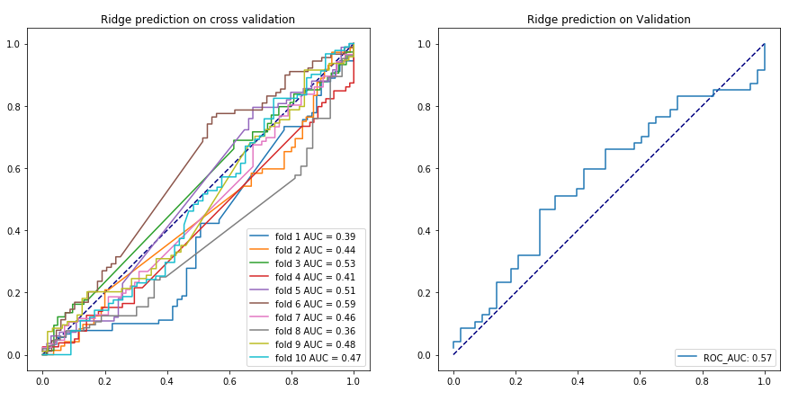
Датасет в серднем показывает ROC-AUC в районе 50%. Все новости, выходящие на данном портале, уже учтены в цене биткойна.

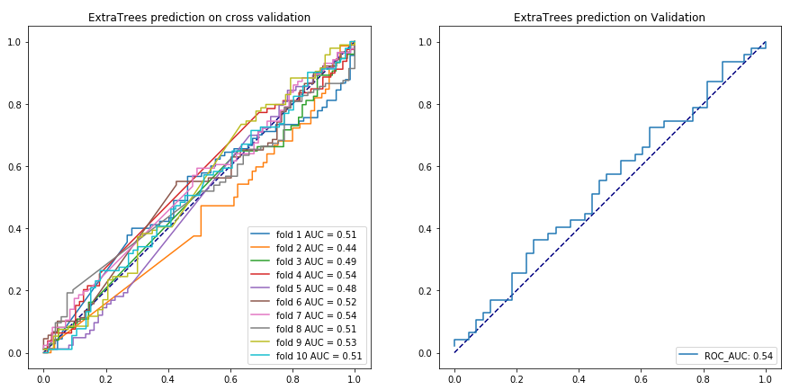




**Моделирование Forbes**

Для метода ExtraTrees на фолдах с 6-ого по 10-ый мы можем наблюдать ROC-AUC выше 0,5. Также это подтверждается на валидации, где мы видим AUC выше 0,5. Это говорит нам о том, что, несмотря на то, что Forbes по большей части состоит из авторских колонок, в целом сайт имеет влияние на криптоактив.





**Bloomberg**

Для Bloomberg анализировалась связь между минутой выхода новости и переменной в следующую минуту. Для обоих моделей – регрессии Ридаж и Extra Trees - мы видим хорошие результаты на кросс валидации и на валидационной выборке. Как и предполагалось изначально, выход новости на сайте Bloomberg влияет на цену криптоактива. При этом свою роль сыграло то, что мы используем минутные данные – более чувствительные к изменению в информационном поле.

