Федеральное государственное автономное

образовательное учреждение высшего образования

«Пермский государственный национальный исследовательский университет» (ПГНИУ)

Региональный институт непрерывного образования (РИНО ПГНИУ)

Цифровая кафедра

Выпускная аттестационная (квалификационная) работа

по курсу профессиональной переподготовки

«Анализ данных и машинное обучение»

**Классификация электорального поведения на основании признаков, описывающих избирательный округ, на выборах США.**

Разработчики проекта:

Неброев Вадим Дмитриевич

Ивушкин Максим Денисович

Волков Данил Анатольевич

Пермь, 2024

Содержание

[ПАСПОРТ ПРОЕКТА 3](#_Toc185769729)

[СОДЕРЖАНИЕ ПРОЕКТА 4](#_Toc185769731)

[Анализ проблемы исследования 4](#_Toc185769732)

[Исходные данные 6](#_Toc185769733)

[Реализация проекта 8](#_Toc185769734)

[Этап 1. Загрузка и предобработка данных 8](#_Toc185769735)

[Этап 2. Разведочный анализ данных 13](#_Toc185769736)

[Этап 3. Объединение данных и создание новых признаков 17](#_Toc185769737)

[Этап 4. Построение модели 18](#_Toc185769738)

[Заключение 29](#_Toc185769739)

[Библиографический список 30](#_Toc185769740)

**ПАСПОРТ ПРОЕКТА**

**Название проекта:**

**Сведения об авторах:** Волков Данил Анатольевич, Неброев Вадим Дмитриевич, Ивушкин Максим Денисович.

**Цель:** Определение электорального поведения на основании различных признаков, описывающих избирательный округ, на президентских выборах США.

**Задачи:**

1. Выполнить анализ предметной области, обосновать проблему и ее актуальность.

2. Осуществить загрузку данных и подготовку их к анализу количественными методами, включая устранение пропущенных значений.

3. Выполнить предварительный анализ данных, в том числе выявление и обработку выбросов, проверку распределения данных на нормальность, корреляционный анализ.

4. Осуществить моделирование зависимости целевого признака методами машинного обучения, в том числе подобрать наилучшую модель, оценить ее качество и выполнить прогнозирование.

5. Выполнить интерпретацию полученных результатов и сделать выводы о достижении цели.

**Краткое описание проекта:**

Требуется проанализировать данные о результатах голосования на выборах в США в 2020 году и определить, возможно ли спрогнозировать победу одного из кандидатов на основании имеющихся факторных переменных методами машинного обучения. Дать интерпретацию полученным результатам и сделать выводы о влиянии различных факторов на результаты.

**Конкретные ожидаемые результаты:**

Построенная модель классификаций.

**СОДЕРЖАНИЕ ПРОЕКТА**

**Анализ предметной области исследования**

На результаты голосования на президентских выборах в США влияет огромное количество различных математических и нематематических показателей. Начиная с экономики и уровня образования, заканчивая дебатами и переменчивым настроением избирателей. Несмотря на это, каждый новый цикл предвыборной гонки ставит избирателей в положение будущей неизвестности, в котором им приходится так или иначе задумываться о том, кто из кандидатов сможет обеспечить именно ту политику, к которой склоняется избиратель. Чаще всего, ключевыми факторами выбора становятся показатели, которые поддаются измерению: экономика, образование, демография и так далее, а значит существует возможность, на основании этих факторов, приближенно измерить настоящие настроения избирателей.

1. Электоральное поведение

Электоральное поведение в США, в основном, зависит от уровня образования, доходов, этнического состава и географии. Так, например, демократы чаще получают поддержку в городах среди образованных избирателей и этнических меньшинств. Республиканцев поддерживают в сельских районах, среди менее образованных слоёв населения и тех, кто работает в традиционных отраслях. В общем, американская процедура голосования крайне усиливает роль "колеблющихся" штатов, где распределение голосов между партиями наиболее сбалансировано.

2. Образование

Уровень образования в округе также позволяет оценить политические предпочтения. Люди с высшим образованием, как правило, чаще поддерживают демократов, так как образование развивает критическое мышление, толерантность и открытость к новым идеям, соответствующим либеральным ценностям. Они заняты в инновационных отраслях, выигрывающих от глобализации и демократических реформ. Напротив же, люди с низким уровнем образования чаще поддерживают республиканцев, поскольку они придерживаются традиционных ценностей и работают в отраслях, уязвимых к глобализации, таких как сельское хозяйство и промышленность.

3. Экономические показатели

Несомненно, экономические показатели, такие как уровень доходов и неравенство, в большей степени влияют на выборы в США. Люди с высоким доходом, занятые в инновационных и интеллектуальных отраслях, чаще поддерживают демократов, которые продвигают социальные реформы и инвестиции в науку. Напротив, избиратели с низким доходом, работающие в традиционных секторах, таких как сельское хозяйство и промышленность, склонны голосовать за республиканцев, обещающих налоговые льготы и протекционистскую политику. Индекс Джини, отражающий неравенство доходов, также может усиливать политическую поляризацию.

4. Демографические данные

Демографическая ситуация также может показывать текущий вектор предпочтений среднего избирателя. Этнические меньшинства, такие как афроамериканцы и латиноамериканцы, чаще поддерживают демократов, в то время как белые избиратели, склоняются к республиканцам. Высокая плотность населения в городах также ассоциируется с поддержкой демократов, так как городские районы часто более либеральны и разнообразны, тогда как сельская местность с низкой плотностью населения чаще поддерживает республиканцев.

**Исходные данные**

Набор данных, используемый в исследовании, содержит долю голосов на президентских выборах в США 2020 года по избирательным округам. Основным преимуществом набора данных является то, что он содержит различную важную статистику, такую как расовый состав округов, медианный и средний доход, неравенство доходов, плотность населения, уровень образования, численность населения и распределение профессиональных направлений.

* Общее количество атрибутов: 34.
* Количество записей: 3143.

Информация об атрибутах:

Первые столбцы содержат информацию об округе и штате.

* Округ
* Штат

В следующих столбцах указаны голоса за 2020 год как в чистом виде, так и в %.

* 2020 Democrat vote
* 2020 Democrat vote %
* 2020 Republican vote
* 2020 Republican vote %
* 2020 other vote
* 2020 other vote %

В следующих столбцах указан уровень образования населения округа.

* Population with less than 9th grade education
* Population with 9th to 12th grade education, no diploma
* High School graduate and equivalent
* Some College, No Degree
* Associates Degree
* Bachelor’s Degree
* Graduate or professional degree

После этого у нас есть информация о доходах и неравенстве доходов в округе.

* Gini Index
* Median income (dollars)
* Mean income (dollars)

Далее у нас есть расовый состав округа. Население округа и плотность населения

* Area in square Km
* Density per square km
* Total Population
* Hispanic or Latino percentage
* White percentage
* Black percentage
* American Indian and Alaska Native percentage
* Asian percentage
* Native Hawaiian and Other Pacific Islander percentage
* Some Other Race percentage
* Two or More Races percentage

В последних столбцах содержится информация о распределении профессий в округе.

* Percentage engaged in Management, business, science, and arts occupations
* Percentage engaged in Service Occupations
* Percentage engaged in Sales and Office
* Percentage engaged in Resources and Construction
* Percentage engaged in Transportation

Необходимо проанализировать данные и определить, возможно ли классифицировать округа по принадлежности к определённой политической партии с помощью методов машинного обучения. В качестве целевой переменной выберем партийную принадлежность округа (например, демократическую или республиканскую).

Гипотеза исследования: принадлежность округа к партии зависит от факторов, таких как, уровень образования, этнический состав, доход, плотность населения, уровень занятости, а также от других демографических характеристик. Эти признаки могут быть использованы для классификации округов и предсказания, к какой партии они будут склоняться на выборах.

Реализация проекта

**Этап 1. Загрузка и предобработка данных**

На этом этапе выполняется загрузка данных из CSV-файла и проводится первичный анализ данных.

Для начала работы необходимо подключить библиотеки.

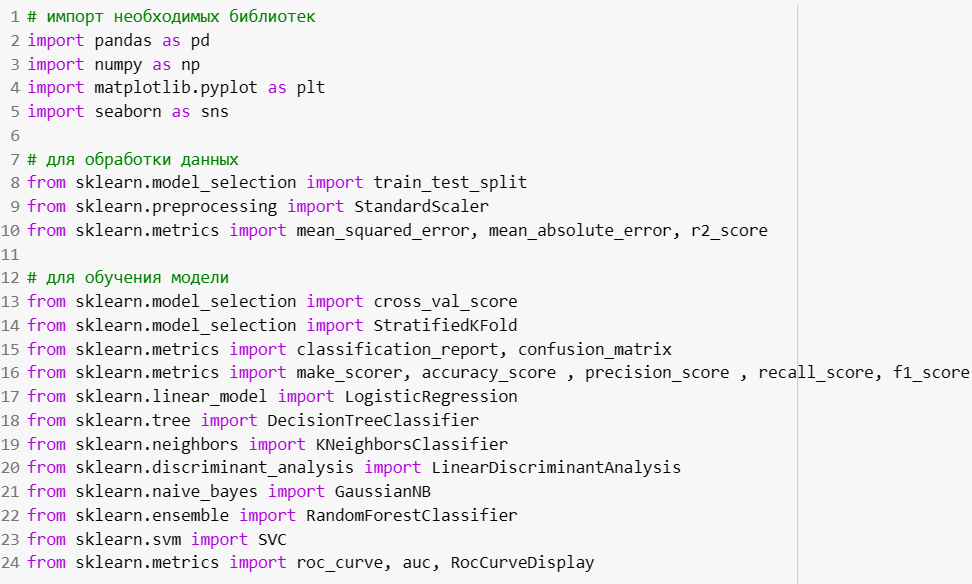


Рисунок 1 Подключение библиотек

Загрузка датасета и вывод данных о нем.

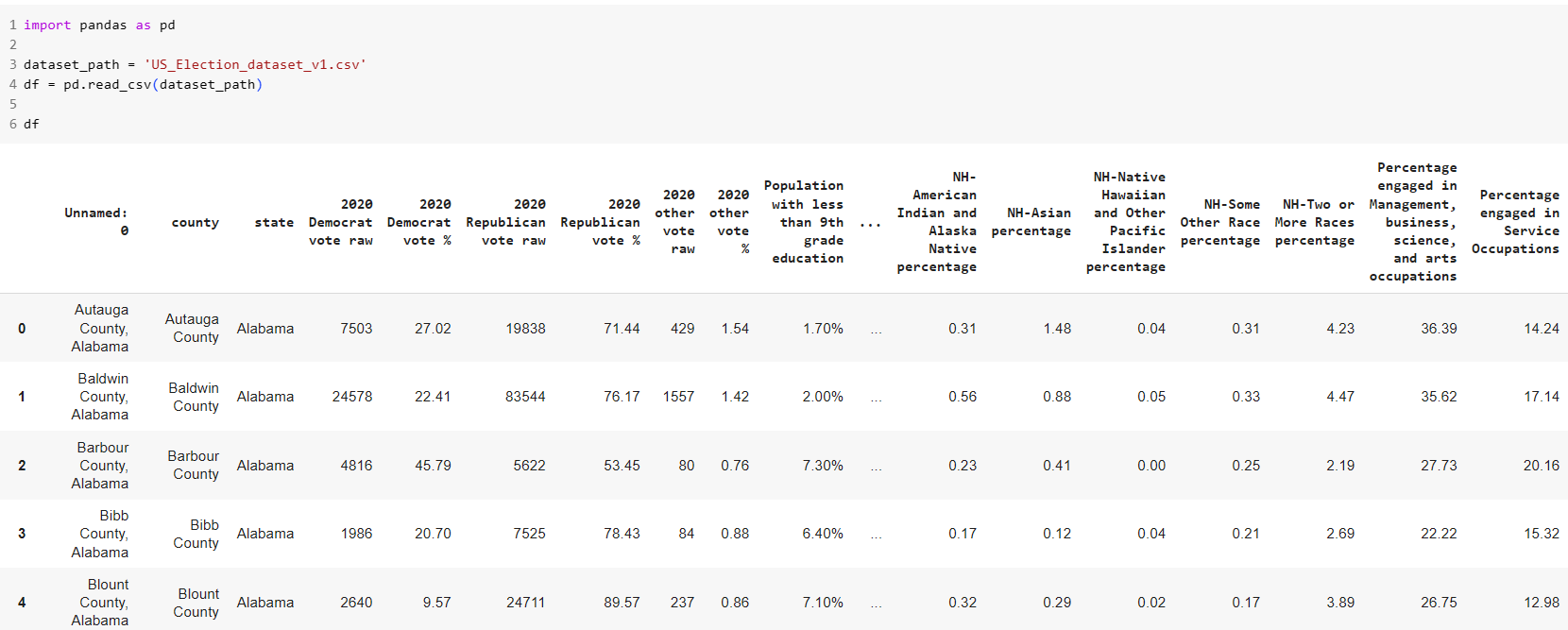


Рисунок 2 Загрузка данных и их вывод

Вывод информации о столбцах датасета и вывод статистической информации.

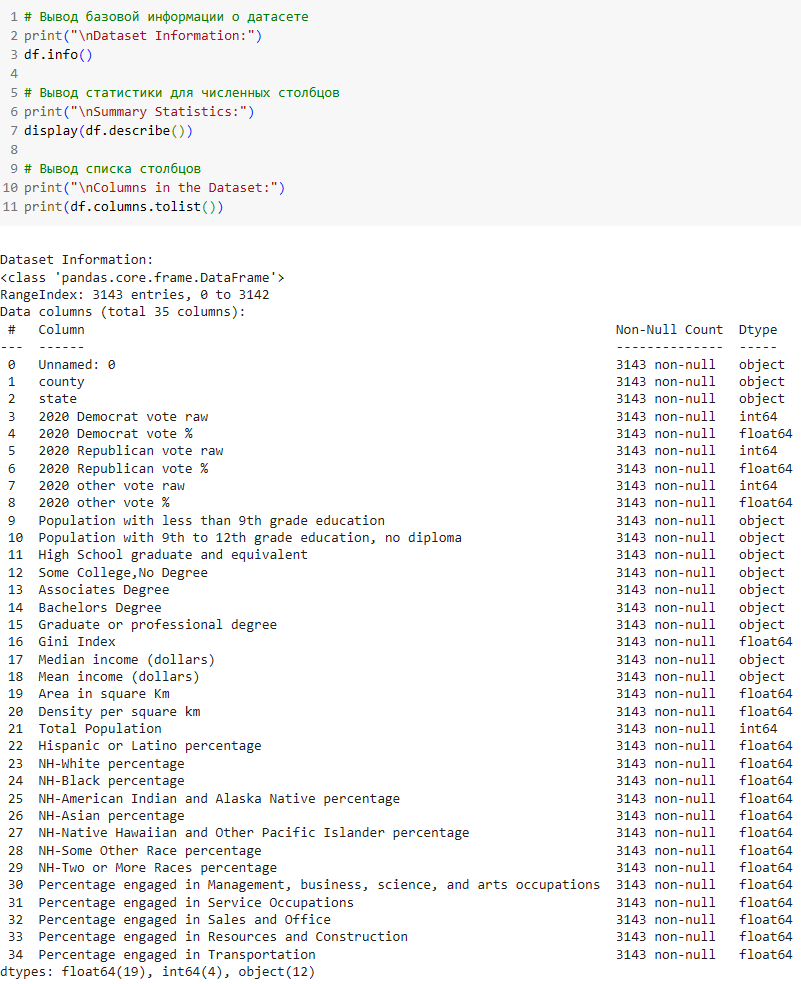


Рисунок 3 Вывод базовой информации

Получение статистических характеристик данных, таких как среднее значение, стандартное отклонение, минимальное и максимальное значения.

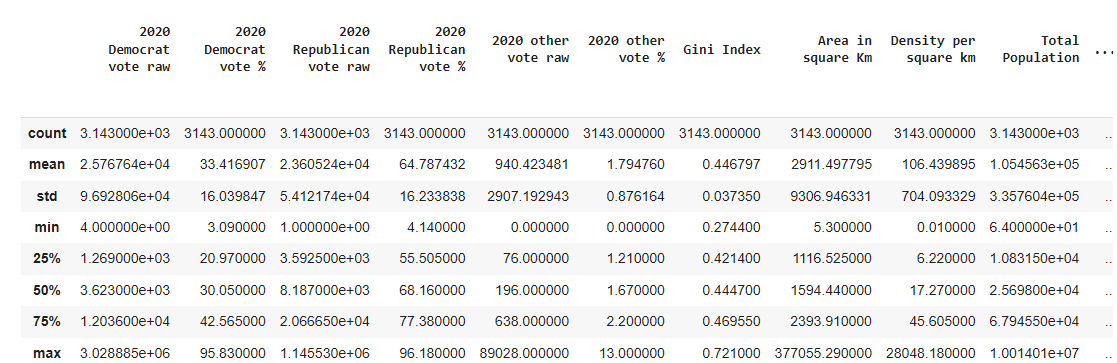


Рисунок 4 Статистические характеристики данных

Очистка и предобработка данных.

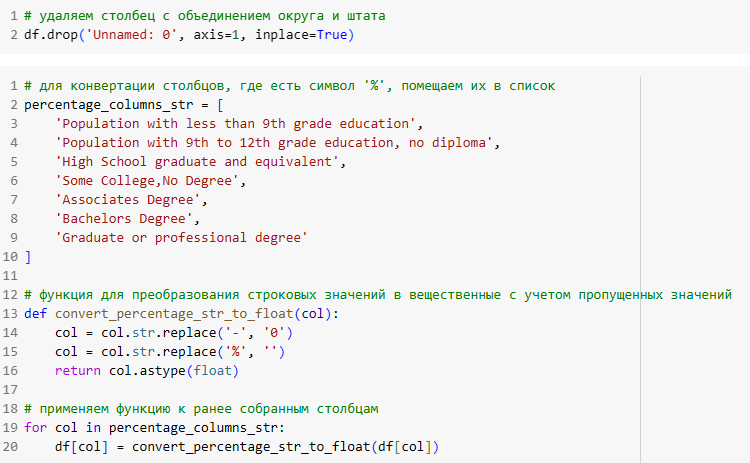


Рисунок 5 Конвертация столбцов

В качестве предварительной очистки удаляем столбец «Unnamed: 0», который является полным названием округа. Следующим шагом будет преобразование строковых признаков, отражающих процентное соотношение в вещественные значения, с учетом возможных пропусков.

Конвертируем столбцы и проверяем остались ли нулевые значения



Рисунок 6 Преобразование столбцов

Ту же процедуры конвертации проводим с оставшимися столбцами, которые, в отличие, от предыдущих, не содержат ненужных символов, а являются строковым представлением вещественных данных.

Для полей, показывающих денежные значения в формате x,yy,zzz, делаем конвертацию с учетом сотен и убираем запятые из строковой записи денежной суммы, а после заполняем пустые значения медианными.

Упрощаем имена столбцов и выводим датасет после преобразований.

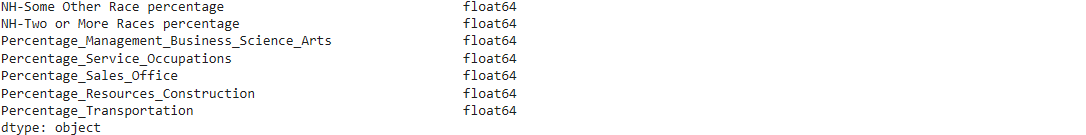
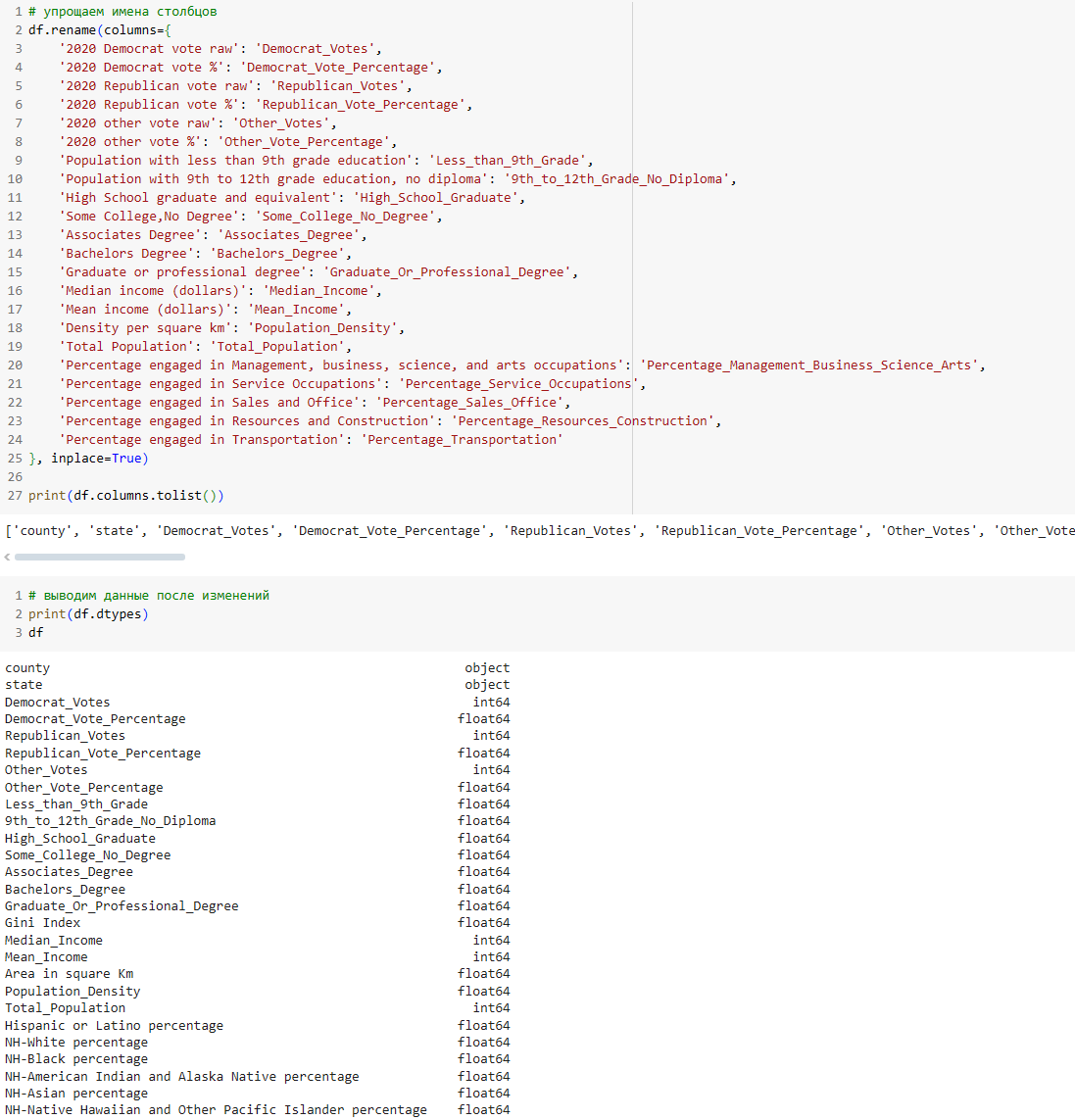


Рисунок 7 Вывод измененных столбцов

**Этап 2. Разведочный анализ данных**

На данном этапе проводится разведочный анализ данных с помощью графиков распределения голосов за демократов и республиканцев. Используются гистограммы с KDE (оценка плотности вероятности), чтобы визуализировать, как распределяются проценты голосов по округам для каждой партии. Это помогает понять, насколько равномерно или неравномерно распределены голоса среди округов для каждой из партий.

Выведем график распределения голосов за демократов.

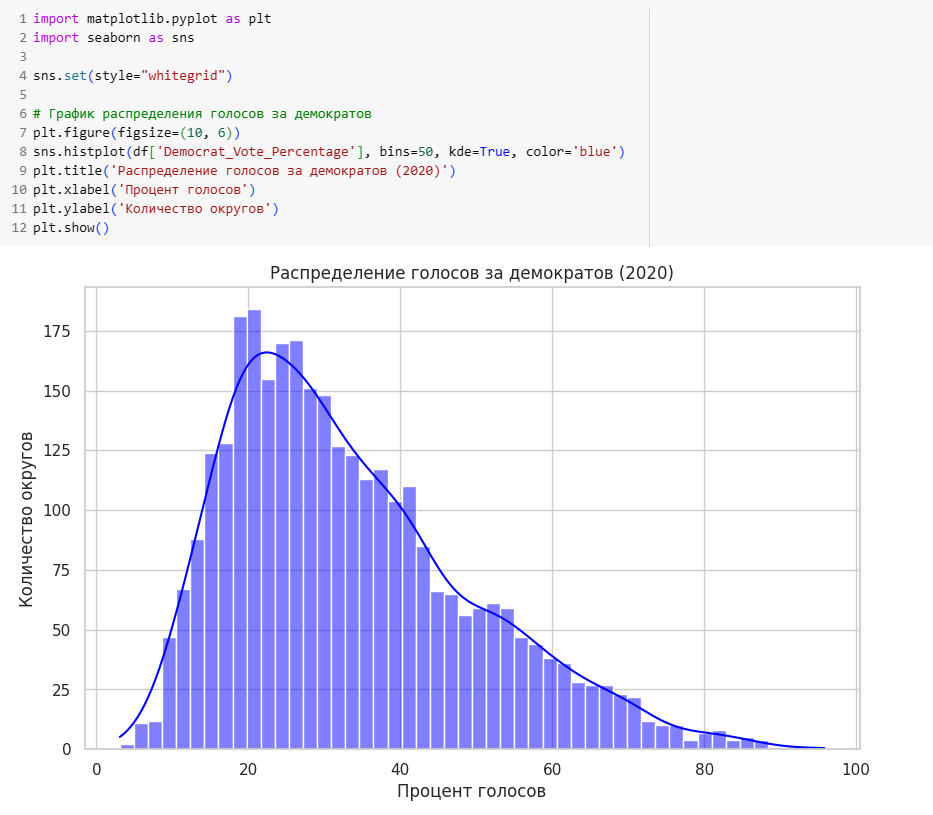


Рисунок 8 Распределение голосов за демократов 2020

Выведем график распределения голосов за республиканцев

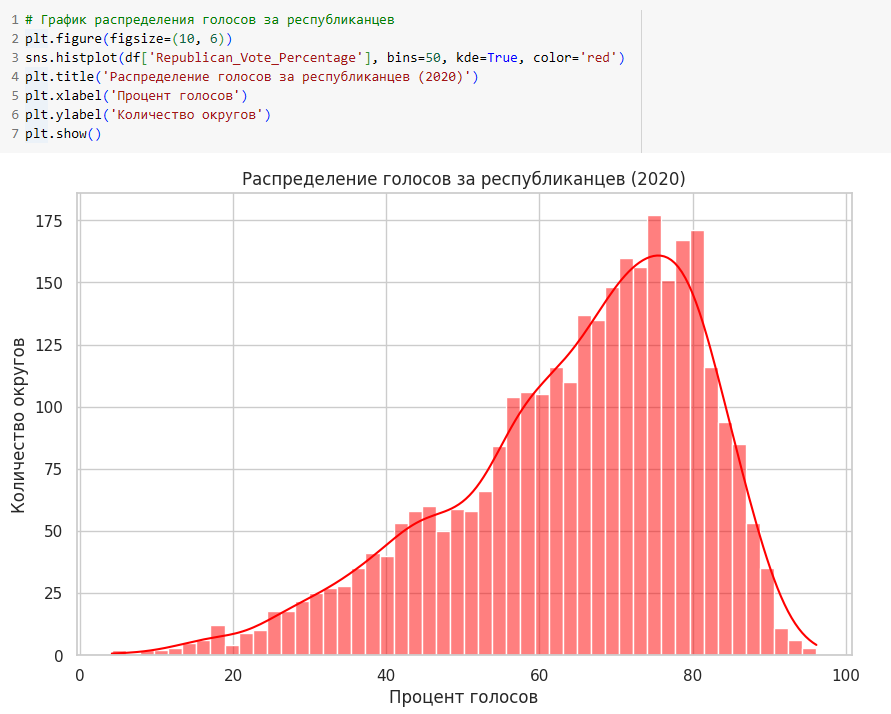


Рисунок 9 Распределение голосов за республиканцев 2020

На данных графиках можно заметить, что в наборе данных существенно преобладают округа, преимущественно голосовавшие за республиканцев.

Затем проводится анализ взаимосвязей между числовыми переменными с помощью матрицы корреляций. Исключаются столбцы, содержащие информацию о округе и штате, и рассчитывается корреляция между оставшимися числовыми переменными. Тепловая карта визуализирует эти связи, где цветовая гамма отражает степень корреляции между переменными, что помогает выявить возможные зависимости и взаимосвязи в данных.

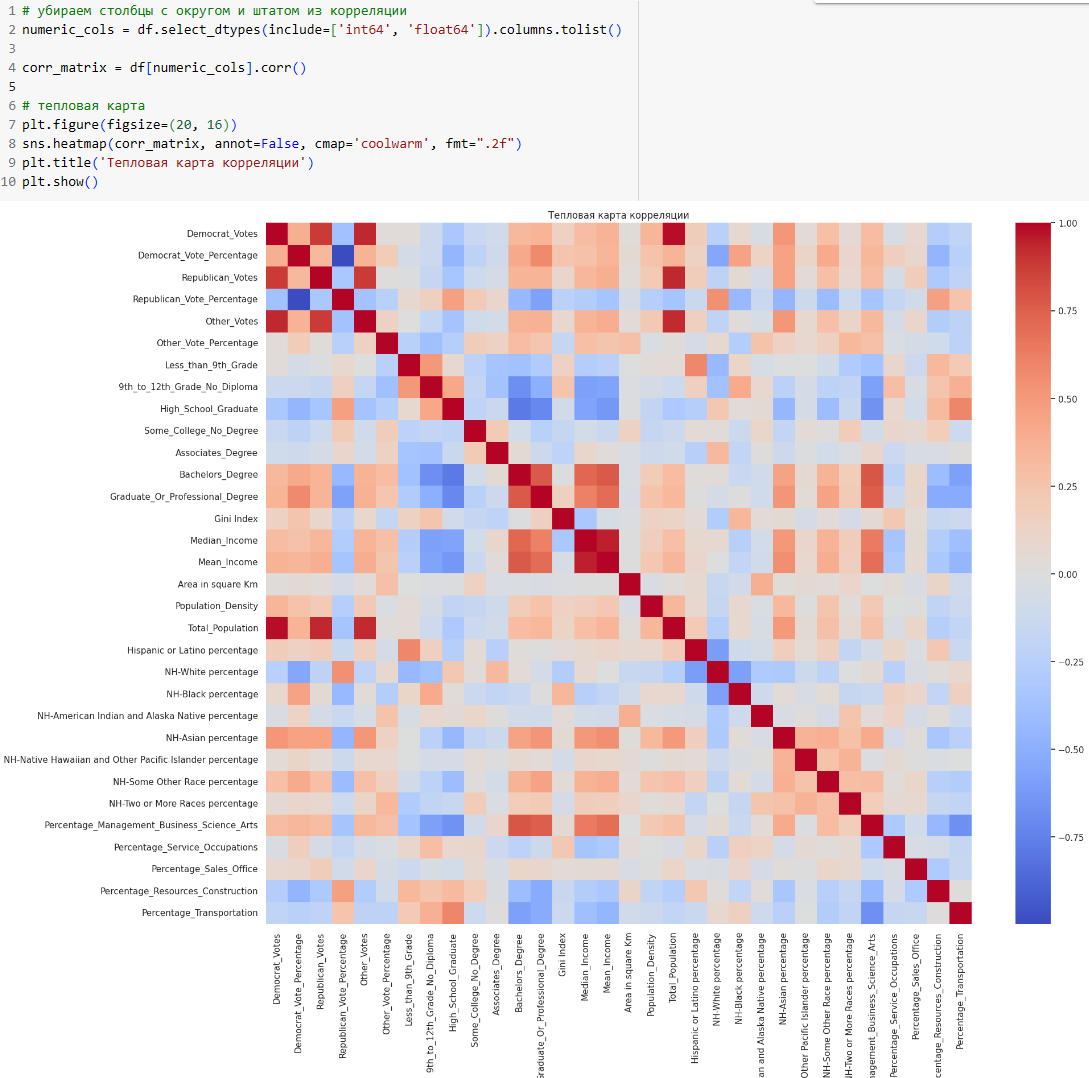


Рисунок 10 Тепловая карта корреляции

На данной тепловой карте представлены зависимости признаков друг от друга. По карте однозначно можно сказать, что в данных нет аномалий.

**Этап 3. Объединение данных и создание новых признаков**

На данном этапе данные были расширены за счёт создания новых признаков для улучшения анализа. Были рассчитаны интегральный индекс образования на основе долей населения с различными уровнями образования, медианный и средний доход на душу населения, а также индекс профессионального разнообразия, отражающий равномерность занятости в разных сферах. Также данные очищены от незначимых столбцов, таких как название округа, штата и площадь территории, чтобы сосредоточиться на ключевых переменных для анализа и моделирования.

Создаем новые признаки.

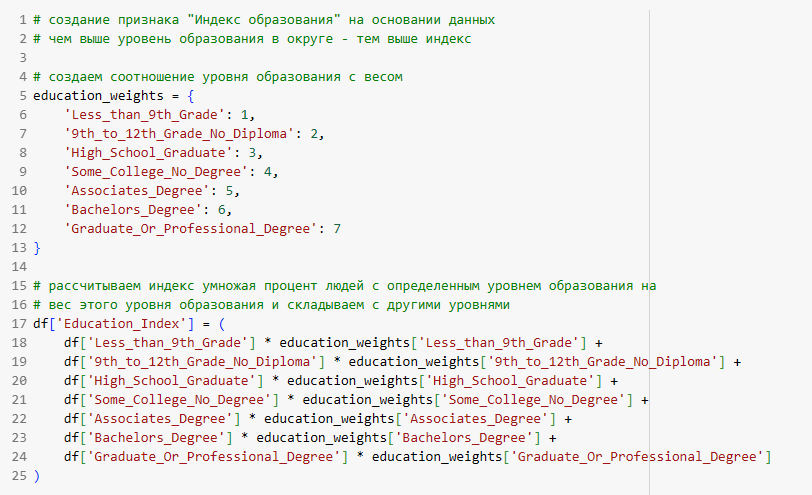


Рисунок 12 Создание новых признаков

Люди с более высоким уровнем образования сильнее влияют на итоговое значение индекса, нежели люди, с более низким уровнем.

Рассчитываем доход на человека в округе, а также создаем признак «Индекс профессионального разнообразия» на основании данных о занятости, чем равномернее распределены различные профессиональные направления - тем выше индекс.



Рисунок 13 Создание новых признаков

**Этап 4. Построение модели**

Подготовка данных к моделированию: переименовываем оставшиеся столбцы и создаем целевой признак на основании результатов голосования. На данном этапе данные подготовлены к моделированию: создан целевой признак State\_Status, который классифицирует округа на основе результата голосования (0 — большинство голосов за демократов, 1 — за республиканцев).

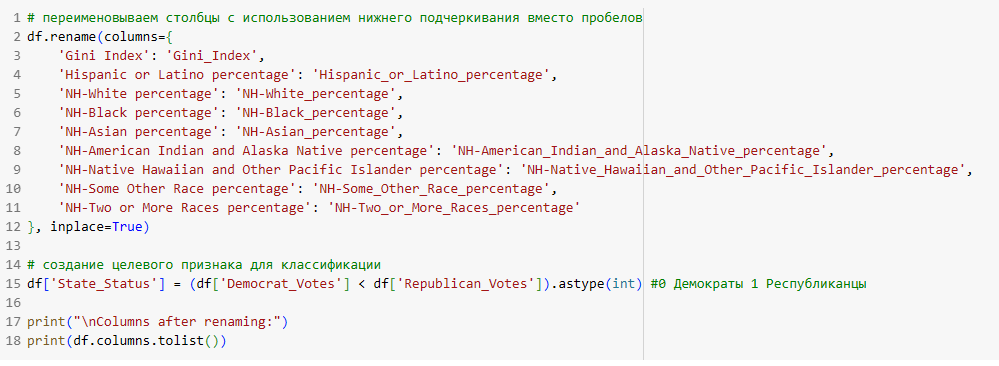


Рисунок 14 Подготовка данных

Выбираем признаки и целевую переменную, а также для балансировки данных используем андерсемплинг.

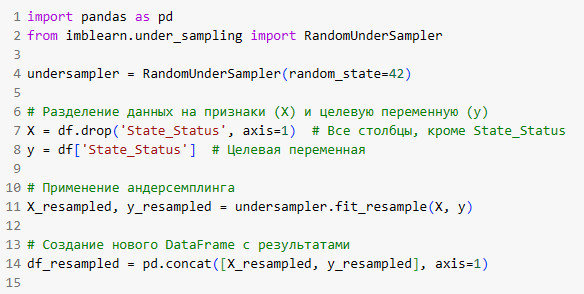


Рисунок 15 Создание нового датафрейма



Рисунок 16 Определение таргета и признаков

Дальнейшие шаги включают:

Разделение данных на обучающую и тестовую выборки, чтобы обучить модель на одной части данных и оценить её эффективность на другой.

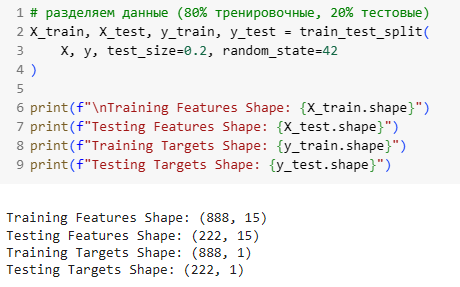


Рисунок 17 Разделение данных

Для обеспечения корректной работы алгоритмов, особенно для моделей, чувствительных к масштабу данных, необходимо нормализовать данные, для этого был использован метод StandardScaler, который используется для изменения размера распределения значений так, чтобы среднее значение наблюдаемых значений было равно нулю, а стандартное отклонение одному.

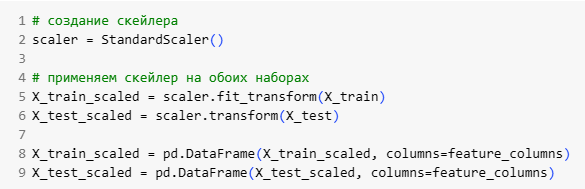


Рисунок 18 Нормализация

Построение нейросетевой модели: Используется Scikit-learn для создания нейронной сети. Настраиваются параметры, такие как количество нейронов, слоёв и количество итераций, после чего модель обучается на подготовленных данных. Далее оценивается её качество на тестовой выборке с использованием метрик классификации.

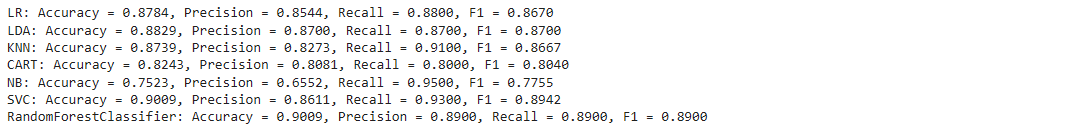


Рисунок 19 Обучение и тестирование моделей

Далее подбираем оптимальные гиперпараметры для модели случайного леса с использованием Grid Search. Задаётся сетка параметров (`param\_grid`), таких как количество деревьев, глубина дерева, минимальное число образцов для разделения и листьев, а также использование бутстрепа. Grid Search перебирает все возможные комбинации параметров, обучает модель для каждой из них и оценивает её с помощью кросс-валидации.

После завершения выбираются лучшие параметры, создаётся модель с этими параметрами, и её качество оценивается на тестовой выборке с помощью метрик точности и классификационного отчёта.

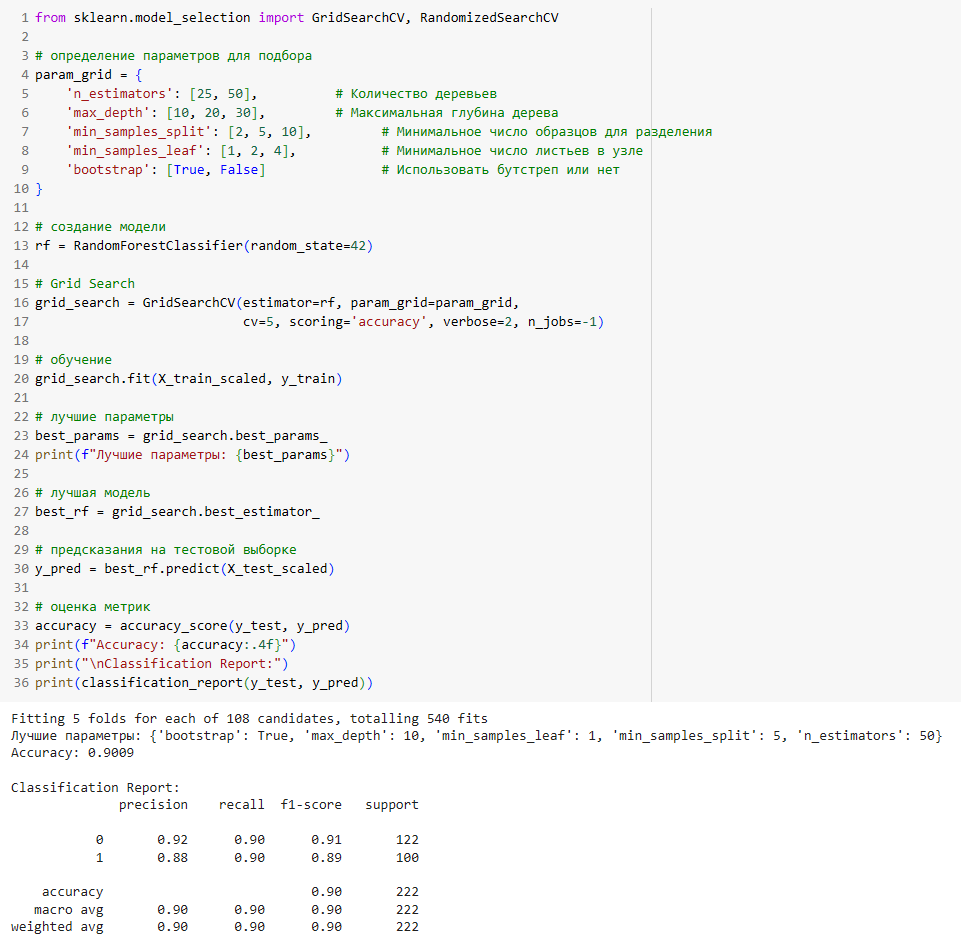


Рисунок 20 Определение параметров

Анализируем важность признаков для модели RandomForestClassifier. Сначала модель обучается на данных X и y, используя заданные гиперпараметры, такие как количество деревьев, глубина и минимальное количество листьев. После обучения извлекается информация о важности признаков с помощью метода feature\_importances\_, который оценивает, насколько каждый признак влияет на качество предсказаний модели.

Затем создаётся DataFrame, где каждому признаку соответствует его оценка важности, а также кумулятивная сумма (cumsum), показывающая общий вклад нескольких наиболее значимых признаков. Это помогает проанализировать, насколько важны отдельные признаки и какая их комбинация покрывает большую часть предсказательной мощности модели.

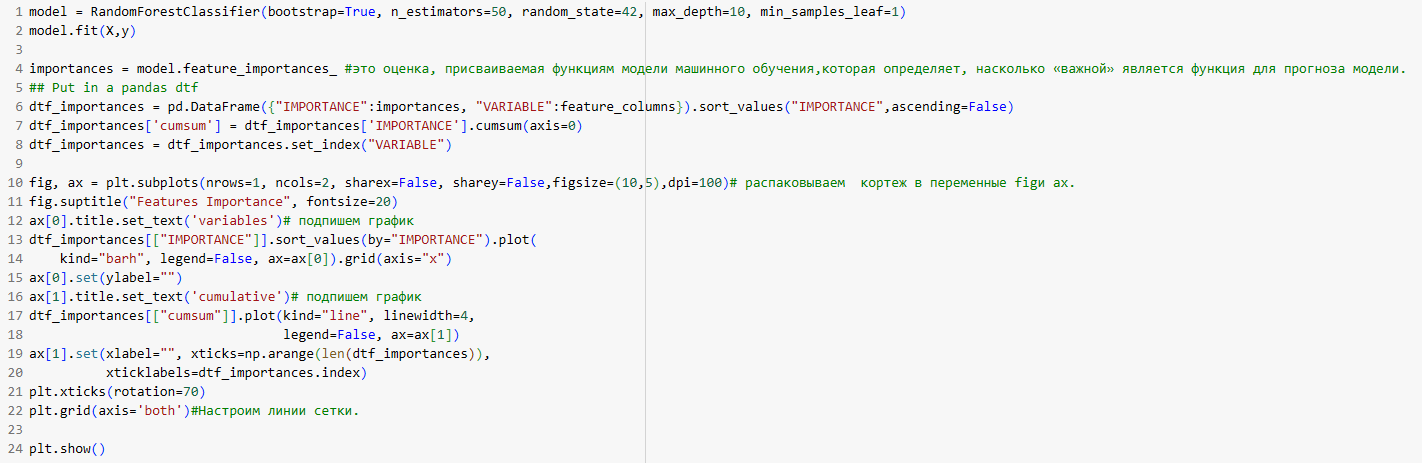


Рисунок 21 Постройка графиков

Для визуализации используются два графика. Первый график представляет горизонтальную столбчатую диаграмму, на которой показаны важности признаков в порядке возрастания, позволяя легко определить наиболее значимые признаки.

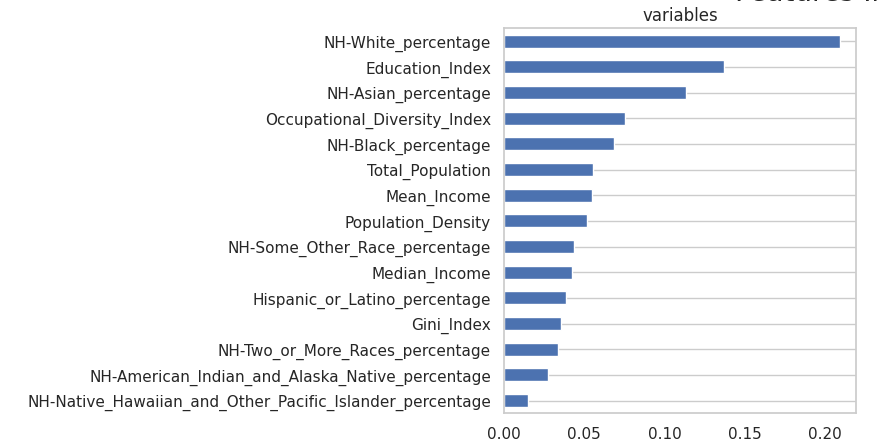


Рисунок 22 Диаграмма важности признаков

На данной диаграмме можно увидеть, что наиболее важные признаки — это индекс образования (Education\_index), а также процент проголосовавших белых людей (NH-White\_percentage).

Второй график — линейный, иллюстрирует кумулятивную важность, что помогает понять, какая часть признаков объясняет основную долю модели. В итоге этот анализ даёт возможность интерпретировать модель, выделить наиболее важные признаки и, при необходимости, сократить их число.

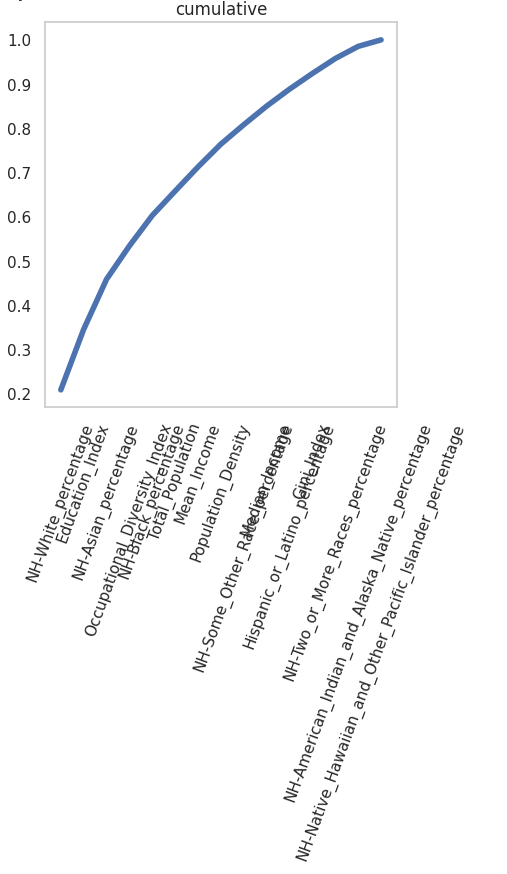


Рисунок 23 График кумулятивной важности

На графике можно заметить, что коренное население континента почти не влияет на выборы, либо влияет на них в пределах погрешности (малонаселенные северные штаты).

После, строим матрицу ошибок, которая позволяет визуально оценить качество классификации. Сначала создаётся матрица ошибок cm с помощью функции confusion\_matrix, которая сравнивает истинные значения целевой переменной и предсказания модели.

Затем матрица визуализируется с использованием тепловой карты (heatmap) библиотеки Seaborn. Элементы матрицы показывают количество верных и ошибочных классификаций для каждой категории.

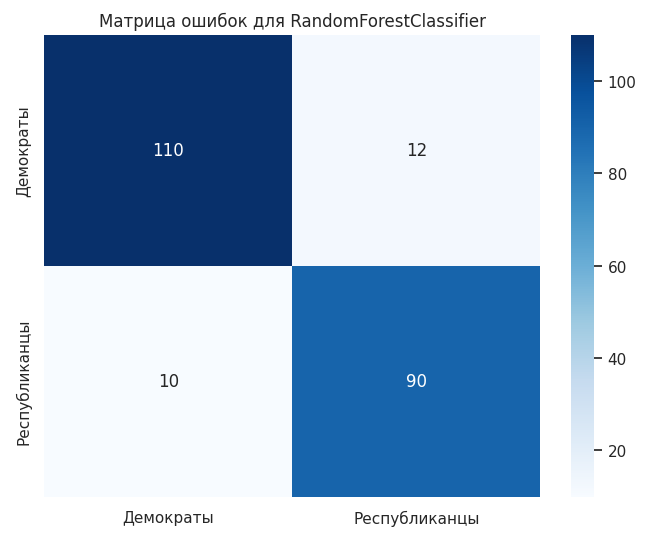


Рисунок 24 Матрица ошибок

На данном рисунке можем увидеть, что количество неправильно предсказанных значений достаточно мало и находиться в пределах 20%.

Далее анализируем качество модели RandomForestClassifier для различных порогов вероятности, которые используются для принятия решения о принадлежности объекта к определённому классу.

Сначала данные делятся на обучающую и тестовую выборки. Затем модель обучается на обучающей выборке, а для тестовой выборки вычисляются вероятности принадлежности к классу.

Цикл перебирает пороги вероятности от 0.1 до 1 с шагом 0.1. Для каждого порога создаётся предсказание: если вероятность больше порога, объект относят к классу "1", иначе — к классу "0". Далее для каждого порога вычисляются метрики качества: accuracy (точность), precision (доля правильных предсказаний для положительного класса), recall (полнота, сколько положительных объектов правильно классифицировано), и f1-score (среднее гармоническое между precision и recall).

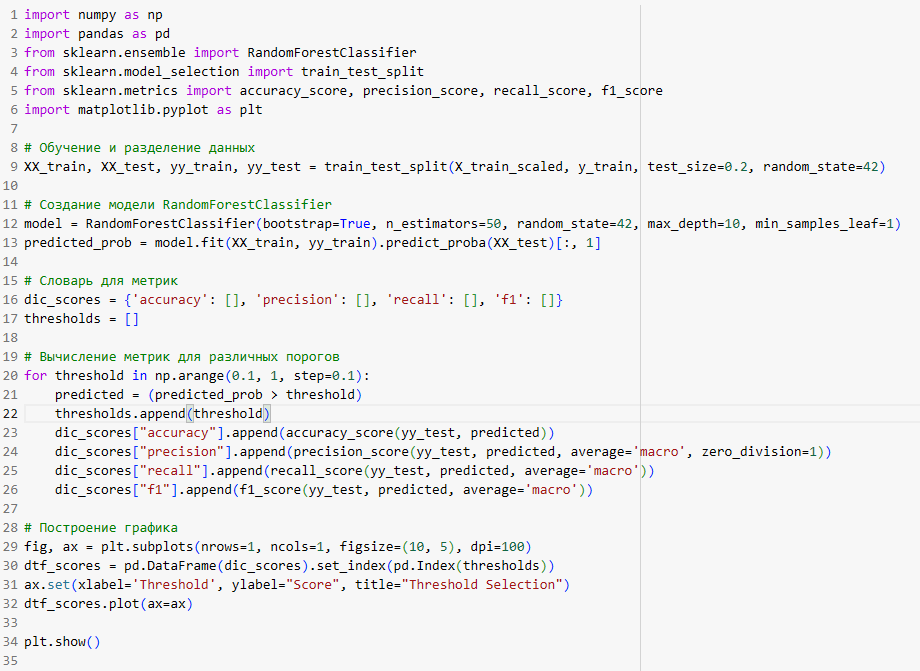


Рисунок 25 Вычисление метрик и построение графика

Результаты всех метрик сохраняются в словарь и затем отображаются на графике. Этот график помогает выбрать оптимальный порог вероятности для балансировки между различными метриками качества модели.

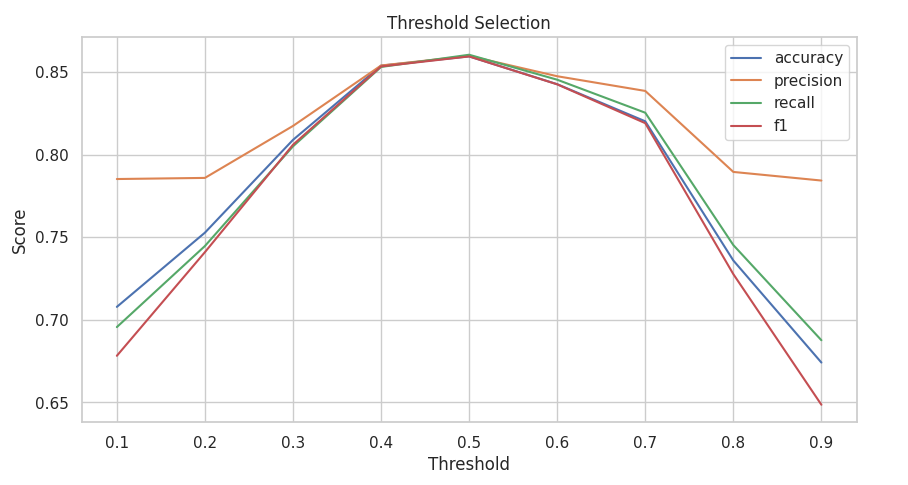


Рисунок 26 График выборки порогов вероятности

По рисунку можно сделать вывод, что оптимальное значение Threshold равно 0.5, так как в данной точке метрики показывают наилучшие значения.

**Заключение**

В ходе исследования мы поставили перед собой цель определить электоральное поведение на основании различных признаков, описывающих избирательный округ, на президентских выборах США.

Для достижения поставленной цели был проведен анализ предметной области, была поставлена проблема, осуществлена загрузка данных и подготовка их к анализу, далее был выполнен предварительный анализ данных, поиск выбросов, проверка распределения данных на сбалансированность, были выбраны наиболее коррелирующие факторные признаки, на основании этого, была подобрана модель и выполнена оценка ее качества.

В целом у модели и данного исследования есть большой потенциал к доработке, но полученные результаты уже можно считать валидными и удовлетворительными в контексте данного исследования.

**Библиографический список**

1. Ruth Igielnik, Scott Keeter, Hannah Hartig «Behind Biden’s 2020 Victory» URL: <https://www.pewresearch.org/politics/2021/06/30/behind-bidens-2020-victory/> (дата обращения 22.12.2024)
2. «Machine Learning in Python» URL: <https://scikit-learn.org/stable/> (дата обращения 22 12.2024)
3. «US Election Dataset» URL: [https://www.kaggle.com/datasets/essarabi/ultimate-us-election-dataset/data](https://www.kaggle.com/datasets/essarabi/ultimate-us-election-dataset/data%20) (дата обращения 22.12.2024)
4. «Learn Data Science Tutorial With Python» URL: <https://www.geeksforgeeks.org/data-science-with-python-tutorial/?ref=shm> (дата обращения 22.12.2024)