Distr.: General

20 октября 2022 г.
14:55:11

Русский

Европейская экономическая комиссия

Конференция европейских статистиков

Группа экспертов по статистике миграции

Женева, Швейцария, 26-28 октября 2022 года
Пункт A предварительной повестки дня
Положительные изменения в использовании административных данных для статистики миграции

Применение машинного обучения для классификации долгосрочных международных мигрантов в Великобритании с помощью административных данных

Записка Национальной статистической службы

Аннотация

Национальная статистическая служба (НСС) трансформирует статистику населения, миграции и социальную статистику, используя сочетание административных данных и данных обследований. Оценка международной миграции с использованием административных данных сопряжена со значительными трудностями, в том числе с получением своевременных оценок. Оценки на основе административных данных характеризуются неизбежной временной задержкой из-за времени, необходимого для сбора и обработки данных, а также взаимодействия со службами, например, в течение 12 месяцев или более, чтобы соответствовать определению долгосрочного мигранта ООН.

НСС разработала оценки миграции на основе административных данных (ОМАД). В настоящее время ОМАД используют подход, основанный на правилах, для классификации долгосрочных международных мигрантов (ДММ). Для завершения классификации целой когорты мигрантов требуются данные за пять лет. Мы предлагаем оценить, как прогностическая модель машинного обучения может обеспечить своевременный прогноз на агрегированном уровне до классификации на основе правил. Мы также хотели бы оценить использование машинного обучения для классификации возвращающихся мигрантов и эмигрантов из Великобритании, а также его способность учитывать значительные изменения в поведении мигрантов, например, во время пандемии коронавируса (COVID-19). Мы разрабатываем эти методы с экспертами в области методологии НСС, используя данные выездных проверок Министерства внутренних

*Подготовили Минцин Ву и Майкл Хокс

дел для подтверждения работоспособности концепции, оценивая точность ранних прогнозов итогового статуса ДММ и то, как эти прогнозы можно использовать в качестве ранних показателей приблизительного количества ДММ для улучшения предварительной оценки миграции. В этой статье мы обсудим ход работ и проблемы в этом проекте. Поскольку работа еще не завершена, мы будем рады получить отзывы о методах работы.

I. Введение

1. Миграция - это один из трех основных компонентов изменений в любом населении. Совершенствование статистики международной миграции в Великобритании является вопросом высокой важности для Национальной статистической службы (НСС). С учетом продолжающегося воздействия пандемии коронавируса (COVID-19) и продолжающихся изменений в миграционной политике после выхода Великобритании из Европейского союза (ЕС), последние оценки чистой международной миграции НСС показывают, что в Великобританию прибыло примерно на 239 000 человек больше, чем уехало, в год, закончившийся в июне 2021 года, главным образом за счет иммиграции из стран, не входящих в ЕС (НСС, 2022a). Для поддержки этих усилий необходимы точные и своевременные оценки количества долгосрочных международных мигрантов (ДММ) в Великобританию и из Великобритании, что позволит судить об объемах и влиянии миграции на население Великобритании для проведения государственной политики и принятия решений.
2. Текущие оценки международной миграции НСС являются экспериментальными и основаны на сочетании административных данных и статистического моделирования. Эти экспериментальные оценки основаны на лучших доступных данных и используют определения на основе правил для классификации ДММ в административных данных. Из-за неизбежных задержек, характерных для административных данных, и времени, необходимого для подтверждения статуса ДММ, для получения полной картины всей когорты мигрантов может потребоваться до 2 лет. Для получения более своевременных оценок долгосрочной международной миграции необходимы новые данные и методы, которые позволят более своевременно классифицировать мигрантов как долгосрочных или краткосрочных.
3. В этом документе описываются предварительные исследования подходов к машинному обучению для классификации ДММ, зарегистрированных в административных данных, наряду с краткосрочными международными мигрантами (КММ) и мигрантами, которые еще не полностью соответствуют критериям, основанным на правилах, которые должны быть определены как КММ или ДММ. Ключевой целью такого метода было бы прогнозирование вероятности того, что мигрант, зарегистрированный в административных данных, будет ДММ до того, как станет известен его истинный краткосрочный или долгосрочный статус. Эти вероятности, спрогнозированные для всех потенциальных ДММ, затем можно было бы использовать для предварительной совокупной оценки ДММ, которая бы включала установленных ДММ (то есть тех, кто соответствует критериям, основанным на правилах) и прогнозируемых ДММ, истинный статус которых еще не известен. Мы используем данные проверок при выезде в ходе пограничного контроля в качестве исходного

набора данных для разработки метода-прототипа только для иммигрантов. В этом документе также будет представленаинформация о наших успехах в изменении характеристик населения и миграции, а также о проблемах, с которыми мы сталкиваемся.

II. Потребность в новых методах

A. Международная миграция меняется

4. В последнее время наблюдается изменение моделей международной миграции в Великобритании, вызванное последствиями пандемии COVID-19, выходом Великобритании из ЕС и введением новой иммиграционной системы в конце 2020 года (Министерство внутренних дел, 2020b).
5. Например, пандемия COVID-19 стала серьезной проблемой для Великобритании и статистики, на которую мы полагаемся. Мы сравнили оценки объемов миграции по данным МОП за тот же период в предыдущие годы с данными о пассажиропотоках, включая воздушные, железнодорожные и паромные перевозки за март 2020 года, и обнаружили, что поездки по всем маршрутам резко сократились. Авиаперевозки резко сократились во втором квартале 2020 года, поскольку были введены строгий карантин и ограничения на поездки. С другой стороны, количество поездок на пароме и по Евротоннелю сократилось не так сильно и, следовательно, привело к соответствующему пропорциональному увеличению поездок с использованием паромов и Евротоннеля (HCC, 2021).
6. В марте 2020 года визовые центры также пришлось закрыть. Наши предыдущие анализы показывают заметное сокращение количества рабочих, учебных и семейных виз, поданных и выданных гражданам стран, не входящих в ЕС, во втором квартале 2020 года (HCC, 2021).
7. После выхода Великобритании из ЕС миграционная политика постоянно менялась. Например, с 1 января 2021 года граждане ЕС, желающие переехать в Великобританию, будут подчиняться тем же правилам, что и граждане остальных стран мира, за исключением граждан Ирландии, которые могут продолжать переезжать в Великобританию без ограничений по отдельным соглашениям. Граждане ЕС, уже проживающие в Великобритании, должны были подать заявку в Системе поселения ЕС, если они хотели продолжать жить в Великобритании.
8. Из-за изменений в международной миграции, подобных описанным выше, существует высокий уровень интереса к пониманию того, как меняются модели миграции и что это означает для общества и экономики. Надежные и своевременные статистические данные о населении и миграции необходимы для разработки политики в ответ на эти изменения, и они лежат в основе множества других статистических данных, которые поддерживают решения и дают пищу для общественного обсуждения. HCC приступила к реализации широкой программы исследований по трансформации статистики миграции, а пандемия COVID-19 показала, что необходимо ускорить обновление данных и методов, включая административные данные и методы моделирования, для получения своевременных показателей международной миграции для удовлетворения потребностей широкого круга пользователей (ONC, 2022a).

B. Статистика НСС трансформируется

9. НСС стремится предоставлять наилучшие сведения о населении и миграции, работая с другими государственными ведомствами и используя ряд новых и существующих источников данных для удовлетворения потребностей наших пользователей. Это становится все более важным в быстро меняющемся политическом и социальном контексте, когда мы знаем, что нашим пользователям нужны более надежные фактические данные для поддержки принятия решений как на национальном, так и на местном уровнях.

10. Текущая статистика населения НСС в значительной степени опирается на данные переписи населения, которая проводится раз в десять лет. Хотя это позволяет получать подробные данные для самых малых географических районов каждые 10 лет, в промежуточные годы уровень детализации снижается, а качество оценок населения сокращается по мере удаления от года переписи.

11. Ранее источником данных о международной миграции в Великобритании был Международный опрос пассажиров (МОП), который фиксировал намерения мигрантов оставаться в Великобритании или за ее пределами в течение следующих 12 месяцев. МОП имеет некоторые ограничения в отношении измерения показателей иммиграции и эмиграции, поскольку это выборочное обследование, и опрашиваются не все мигранты, въезжающие в Великобританию или выезжающие из страны. Кроме того, МОП не учитывает всех просителей убежища, которые могут въезжать в Великобританию или покидать ее. МОП основан на первоначальном намерении относительно периода пребывания и не принимает во внимание изменяющиеся намерения пассажиров. Кроме того, он не учитывает тех, кто пересекает сухопутную границу между Великобританией (Северная Ирландия) и Республикой Ирландия. Существует широкий консенсус в отношении того, что МОП вышел за рамки своей первоначальной цели, и теперь НСС рассматривает дополнительные данные и методы для преобразования статистики миграции (НСС, 2018).

12. Кроме того, МОП был приостановлен в марте 2020 года из-за пандемии COVID-19. Несмотря на то, что МОП был возобновлен в 2021 году, он был больше ориентирован на статистику путешествий и туризма. Из-за ограничений по времени и местоположению, что приводит к небольшому количеству контактов, новые данные МОП могут скрывать модели миграции, в частности особенности (например, сезонность) во временных рядах (НСС, 2022b). Такие изменения в МОП подчеркивают необходимость использования административных источников данных и новых методов для заполнения пробелов в фактических данных (НСС, 2022a).

III. Классификация ДММ в данных проверки при выезде

A. Данные проверки при выезде

13. Исходные административные данные, которые мы используем для разработки нашего прототипа классификатора машинного обучения, получены из системы сопоставления данных и аналитического потенциала - системы Анализа начального статуса (АНС), созданной в рамках Программы проверок при

выезде Министерства внутренних дел Великобритании. Мы используем данные из системы АНС Министерства внутренних дел, которая объединяет информацию о визах и поездках, чтобы связать между собой поездки одного человека в страну и из страны. Этот набор данных известен как набор данных проверок при выезде (Министерство внутренних дел, 2020a).

14. Эти данные включают предварительную информацию о пассажирах (ПИП) и информацию о проездных документах (ИПД). ПИП — это данные о пассажирах, представляемые до поездки для большинства регулярных авиаперевозок, а ИПД — это данные о пассажирах, собираемые в пункте отправления для других видов транспорта.
15. Есть также данные из систем обработки дел, связанных с рассмотрением дел (за пределами страны) при подаче заявления на получение въездной визы и работой с делами (внутри страны), например о продлении вида на жительство, а также биометрические данные, представленные до подачи заявления на получение визы (у нас нет доступа к такой информации), и данные проверки паспортов, собранные при въезде в Великобританию. Эти данные помогают устранить любые пробелы в охвате ПИП для въезжающих в страну.
16. Эти данные сопоставляются, чтобы создать «идентификатор» и определить текущий статус соответствия для разных лиц. Каждому лицу в системе АНС присваивается уникальный идентификатор, который состоит из биографических данных (таких как имя, номер паспорта или проездного документа, дата рождения, гражданство и пол) и связанных событий (таких как поездки в страну или из страны или периоды предоставленного отпуска). Соответствующий набор данных называется «Анализом начального статуса».
17. Как и в случае всех крупных сложных наборов данных, основанных на административных источников, полученные данные могут быть не всегда полными и абсолютно точными. Существует также ряд случаев, при которых отъезд путешественника может быть законно не зарегистрирован системой, например в случае Единого иммиграционного пространства (CTA) или когда путешественник, к сожалению, умирает в Великобритании.
18. Для получения дополнительной информации о данных проверок при выезде см. документацию Министерства внутренних дел (Министерство внутренних дел, 2020a).

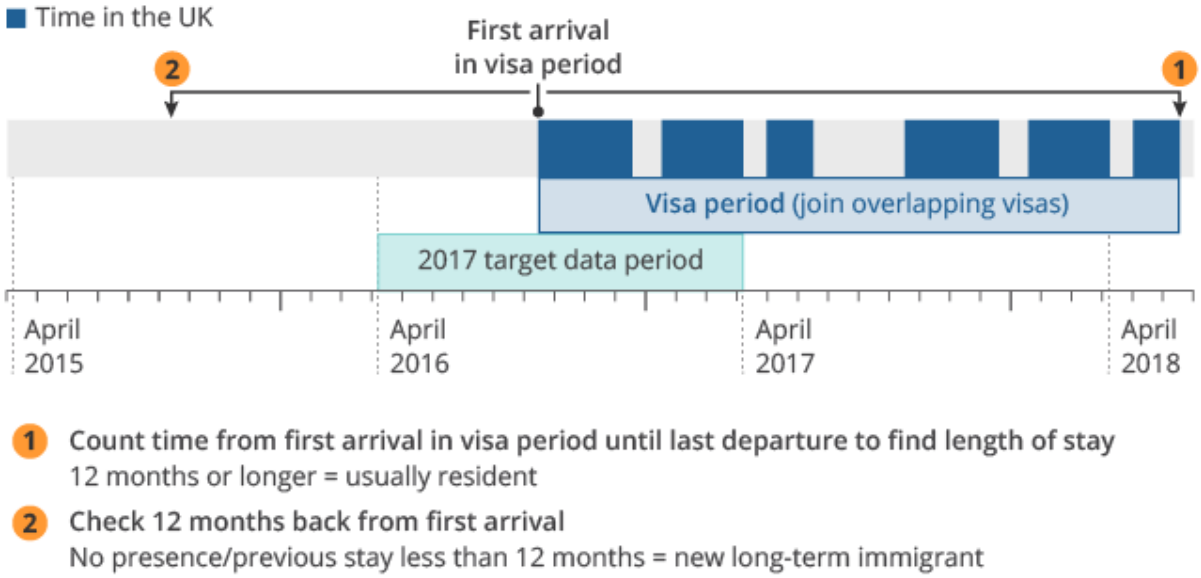
B. Текущий метод классификации

19. Текущий метод классификации ДММ, не являющихся гражданами ЕС, - это статический детерминистский подход, основанный на фактических результатах действий в прошлом по данным проверок при выезде. Мы принимаем определение ООН для этого основанного на правилах процесса для ДММ: «Лицо, которое переезжает в страну, отличную от страны его/её обычного проживания на период не менее 1 года (12 месяцев), и таким образом страна назначения определенно становится новой страной (местом) его/ее обычного проживания. С точки зрения страны выезда человек будет долгосрочным эмигрантом, а с точки зрения страны назначения – долгосрочным иммигрантом» (ООН, 1998).
20. Время пребывания мигранта в Великобритании рассчитывается на основе первой даты въезда и последней даты выезда в период действия визы (если визы пересекаются, они объединяются). Если время пребывания в Великобритании превышает 12 месяцев, каждый период пребывания за

пределами Великобритании меньше 12 месяцев и этот человек не был в стране менее чем за 12 месяцев до даты первого прибытия, то мигрант классифицируется как новый ДММ. Эта показано на Рисунке 1 (HCC, 2020). Группе экспертов ЕЭК ООН по статистике миграции был представлен отдельный документ по текущему методу оценки: «Использование административных данных для получения своевременных оценок миграции в Великобритании».

Рисунок 1

Правила маркировки ДММ в данных проверок при выезде



Источник: HCC (2020)

C. Польза контролируемого машинного обучения для классификации ДММ в Великобритании

1. Актуальность

21. Из-за временных задержек в текущем методе классификации мы предлагаем оценить, как прогностическая модель контролируемого машинного обучения может обеспечить своевременный прогноз для отдельных иммигрантов, зарегистрированных в административных данных, до того, как можно будет применить классификацию на основе правил. Затем эти отдельные прогнозы можно было бы агрегировать, что позволит дать предварительные оценки ДММ.

22. Контролируемое обучение начинается с данных, помеченных вручную. Мы используем текущий метод классификации на основе правил для маркировки ДММ, как описано выше (Рис. 1). Затем помеченные исторические данные используются для обучения модели контролируемого обучения, чтобы можно было делать прогнозы о статусе ДММ мигрантов, которые прибыли настолько недавно, что невозможно применить определения данных. HCC изучает возможность внедрения альтернативных определений, которые будут

применяться в данных проверок при выезде, и мы можем изучить их использование для создания обучающих меток в будущем (HCC, 2020).

23. Контролируемые методы устанавливают взаимосвязи между входными признаками, описывающими мигрантов (например, тип визы, дата окончания действия визы, количество дней с момента прибытия), и их помеченным результатом: ДММ или не ДММ. Затем обученные модели предсказывают вероятность того, что новые мигранты станут ДММ, на основе их входных признаков. Таким образом, модели машинного обучения потенциально могут делать прогнозы на основе индивидуальных характеристик без необходимости ждать данных не менее 12 месяцев, чтобы применить определения данных на основе правил.
24. Использование контролируемого машинного обучения имеет то дополнительное преимущество, что не нужно разрабатывать и поддерживать новые предположения и категории данных на основе правил для недавних иммигрантов для достижения цели классификации, прежде чем мы получим подтверждение на основе дат прибытия и отбытия. В принципе, там, где мы могли бы попытаться вручную определить определенные категории мигрантов с определенными характеристиками как более или менее вероятных ДММ, при изучении взаимосвязей между входными признаками и результатом, модели будут аппроксимировать функцию, которая использует доступную информацию для формирования этого прогноза, то есть модели могут позволить делать прогнозы без явной конкретизации человека (Arthur Samuel, 1959). Модели также могут адаптироваться к изменениям в поведении, если они со временем обновляются (то есть переобучаются) с помощью новых помеченных данных, которые фиксируют эти изменения в поведении.
25. Прогнозы на основе моделей машинного обучения будут подвержены ошибкам и неопределенности, и вполне вероятно, что эти ошибки и неопределенность будут больше для недавних иммигрантов, но они могут послужить основой для предварительных оценок ДММ с возрастающей точностью в течение жизни когорты (до тех пор, пока они могут быть подтверждены с помощью подхода, основанного на правилах). Их также можно использовать в качестве индикаторов для изучения тенденций и раннего планирования.

2. Относительное взвешивание и калибровка

26. Калибровка - это важный аспект обучения классификаторов в машинном обучении. Это дает представление о неопределенности модели, которое впоследствии может быть сообщено конечным пользователям или использовано при дальнейшей обработке выходных данных модели.
27. С технической точки зрения цель классификации состоит в том, чтобы присвоить предсказанные классы непомеченным данным. Однако полезно также учитывать вероятности, предсказанные моделью, лежащие в основе этих предсказанных классов. Такие вероятности можно интерпретировать как относительный вес, и они полезны для анализа недостатков модели и, возможно, для представления неопределенности конечным пользователям. Использование относительной вероятности также может быть полезным, если это приводят к снижению систематической ошибки в предварительных оценках ДММ по сравнению с классификацией (Zhang, 2020).
28. В машинном обучении модель считается хорошо откалиброванной, если вероятности прогнозирования, выдаваемые моделью, соответствуют

фактической вероятности того, что модель верна на этой выборке. Например, если модель предсказывает, что выборка относится к положительному классу с оценкой вероятности 0,8, можно ожидать, что модель будет верна в этом прогнозе примерно в 80% случаев.

29. Калибровку модели можно визуализировать, построив калибровочную кривую — разделив прогнозы модели на дискретные интервалы на основе вероятности прогноза и отобразив пропорции каждого правильно предсказанного интервала в сравнении со средней вероятностью прогноза каждого интервала. Таким образом, идеально откалиброванный классификатор будет отображаться как прямая линия от (0,0) до (1,1).

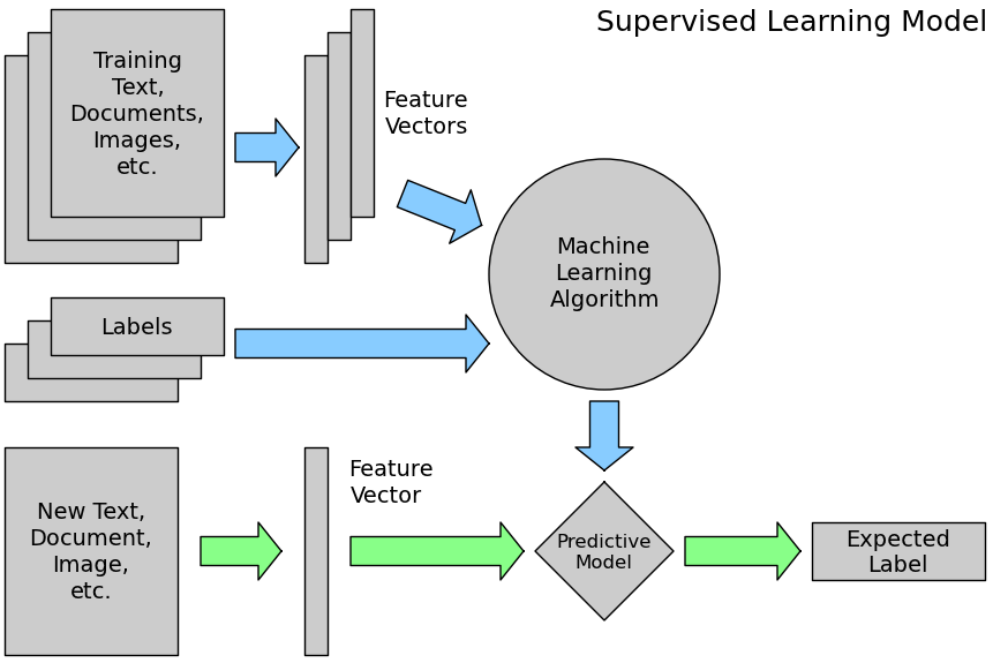
IV. Подход к машинному обучению

30. Этот проект будет использовать данные выездных проверок для учета мигрантов, не являющихся гражданами ЕС, для разработки методов подтверждения концепции. В ожидании результатов этого первоначального исследования в будущей работе будет также рассмотрена аналогичная задача с использованием других источников административных данных, таких как База данных регистрации и взаимодействий населения (RAPID) Департамента труда и пенсий (DWP) для граждан ЕС.

31. Классификация использует обучение с учителем, которое требует маркировки как положительных, так и отрицательных примеров при обучении модели. В нашем случае положительной меткой будет ДММ. Машина использует информацию, полученную при изучении данных ДММ и не-ДММ, для рассмотрения особенностей новых данных, чтобы помочь прогнозировать вероятность того, что мигранты станут ДММ. Подход машинного обучения позволяет нам прогнозировать результаты статусов миграции до того, как подтверждение достигнет конца 12 месяцев. Упрощенный процесс контролируемого обучения показан на Рисунке 2 ниже.

Рисунок 2

Обзор контролируемого обучения



Источник: Pedregosa et al. (2011)

A. Практические подходы к разработке

32. Мы следуем руководству по этике данных Статистического управления Великобритании и тесно сотрудничаем с командой по этике данных HCC и завершили оценку этики для этого проекта. Мы получаем доступ к личной информации только в безопасной среде разработки HCC, и только совокупные результаты выводятся из этой среды при условии одобрения сотрудников по контролю за раскрытием информации.

    В моделировании не используются персональные данные (ПД).

33. Мы следуем рекомендациям Функционального стандарта в области анализа правительства Великобритании и рекомендациям по обеспечению качества государственных моделей (AF, 2021). Мы записываем наши предположения и решения на протяжении всего проекта, чтобы обеспечить качество разработки, анализ проводит группа обеспечения качества, а также более независимые группы экспертов. Мы также используем структуру обеспечения качества для управления нашим воспроизводимым написанием кода и применяем воспроизводимые конвейеры обработки запросов, чтобы избежать человеческих ошибок (GSS, 2022).

B. Базовое население

34. Для апробации концепции модель ориентирована на долгосрочную иммиграцию граждан стран, не входящих в ЕС.

35. Мы начинаем только с иммиграционного компонента, ограничивая наше исследование теми, кто посещал страну до пандемии COVID-19 (апрель 2015 года — апрель 2019 года). Это связано с тем, что на их долю приходится

большая часть данных, и это позволит нам разработать базовые методы для периода, когда Великобритания вышла из ЕС. И только после этого мы займемся последующими изменениями в поведении мигрантов, вызванными пандемией COVID-19.

36. В ожидании разработки первого пилотного метода мы также будем исследовать прогнозирование статуса ДММ для иммигрантов без истории посещений и эмигрантов.

C. Обучение и тестирование

37. Выборка мигрантов, зарегистрированная в данных проверок при выезде, помечена как ДММ или не ДММ на основе определения ДММ ООН, как описано выше. Набор алгоритмов контролируемого обучения был протестирован на выборках ковариат и меток в данных, относящихся к гражданам стран, не входящих в ЕС, чтобы изучить закономерности в данных, прежде чем делать прогнозы, исходя из контрольных данных. В настоящее время мы случайным образом разделяем наш набор данных на обучающий набор, используемый для подбора модели, и тестовый набор, используемый для ее оценки, но мы также рассматриваем возможность стратификации данных таким образом, чтобы ДММ и не-ДММ были в равной степени представлены в обучающем наборе. Кроме того, мы рассматриваем возможность разделения набора данных на обучающие и тестовые наборы по времени (то есть Обучение модели на всех людях, которые взаимодействовали с системой Проверок при выезде до определенной даты, и тестирование на тех, кто взаимодействовал после нее), чтобы лучше оценить, насколько хорошо модель, обученная на исторических данных, работает с будущими данными.

D. Алгоритмы

38. Поскольку ни один алгоритм машинного обучения не будет полезен для всех случаев использования, мы тестируем три алгоритма обучения с учителем, чтобы увидеть, какой алгоритм может обеспечить наилучшие результаты и интерпретацию.
39. Мы рассматриваем классификацию ДММ как проблему бинарной классификации, чтобы предсказать, какие мигранты станут ДММ, и первоначально рассматриваем такие варианты как логистическая регрессия (LR), случайный лес (RF) и XGBoost (XGB). Классификацию ДММ также можно интерпретировать как проблему анализа выживаемости с цензурированными данными, и мы также рассмотрим, насколько нам походят альтернативные методы.

E. Оптимизация модели и оценка производительности

40. Мы намерены оптимизировать обученные модели машинного обучения путем итеративной настройки гиперпараметров с использованием метода автоматической оптимизации гиперпараметров, такого как поиск по сетке, при этом в качестве критерия оценки, который необходимо максимально увеличить в этом процессе, выбрана точность (количество правильных прогнозов, деленное на количество выборок).

41. Следует отметить, что точность может быть плохим критерием использования при наличии значительной несбалансированности классов в обучающих данных, поскольку это может привести к результатам, которые выглядят впечатляюще, хотя на самом деле не являются обученным классификатором — например, если 90% выборок в наборе данных относятся к положительному классу, тогда базовая модель, которая только предсказывает выборки как принадлежащие к этому классу, будет иметь показатель точности 0,9, что делает ее интуитивно похожей на обученную модель.
42. По итогам исследования мы пришли к выводу, что точность будет полезным критерием для нашего проекта, поскольку классы классификации не слишком несбалансированные, и как положительный, так и отрицательный класс имеют почти одинаковое значение.
43. Кроме того, мы будем записывать Recall, Precision и F1-меру каждой модели, чтобы получить более полное представление о ее производительности, но эти метрики не были выбраны для прямой оптимизации, поскольку они, как правило, более полезны в случаях когда интерес представляет только положительный класс в задаче классификации или когда существуют различные затраты, связанные с неправильной классификацией членов положительного или отрицательного класса. Точно так же будет фиксироваться площадь под кривой ошибок (ROC-AUC; площадь под кривой в координатах true positive rate (доля истинно положительных классификаций в общем числе положительных наблюдений) и false positive rate (доля ложно положительных классификаций) для различных пороговых значений модели), что позволит оценить способность модели различать классы.
44. Мы также рассмотрим возможность использования логарифмических потерь модели в качестве метрики для оптимизации, так как это могло бы повысить вероятности прогнозирования, связанные с каждым предсказанием, что было бы полезно при использовании вероятностного взвешивания для определения иммиграционной статистики (вместо подсчета количества людей, прогнозируемых в каждой категории). Однако в данном случае необходимы дальнейшие изыскания, чтобы определить, будет ли такой вариант полезен.

F. Выбор признаков

45. Данные проверок при выезде дают много информации, но не вся эта информация конкретно полезна для прогнозирования того, является ли человек долгосрочным мигрантом или нет. Таким образом, включение каждого доступного нам признака может привести к тому, что модели будут избыточно обучены аномалиям в статистических данных в обучающих данных, что приведет к ухудшению производительности. Сначала мы строим базовые модели, используя небольшое количество атрибутов из данных и сконструированных признаков, и по мере необходимости будем заниматься более сложным конструированием признаков. Чтобы оценить важность каждого признака, мы будем использовать показатели значимости признаков, выдаваемые моделями — средний выигрыш в чистоте от разделения по этому признаку для моделей XGBoost и случайный лес, а также коэффициенты для каждого признака в моделях логистической регрессии. Затем мы сможем вручную проверять признаки, чтобы идентифицировать те из них, которые практически не обладают прогностической силой, а также использовать методы рекурсивного исключения для итеративного обучения моделей с меньшим

количеством признаков, удаляя признаки с наименьшей прогностической силой.

G. Мониторинг и переобучение

46. Производительность модели машинного обучения необходимо постоянно контролировать, чтобы обнаружить возможный сдвиг модели, когда производительность модели ухудшается со временем, особенно при значительных изменениях данных, закономерностей и концепций, используемых в модели, таких как изменения в определениях визы или моделях миграции. В таких случаях может потребоваться переобучение моделей с течением времени, если это необходимо. Нам придется это учитывать, учитывая изменяющийся характер миграционного поведения, о котором мы говорили выше.

H. Приобретенный опыт

47. В настоящее время мы изучаем производительность первых базовых моделей, которые мы обучили, и надеемся, что сможем опубликовать первоначальные результаты в отчете о ходе работ позже в этом году.
48. Один из уроков, которые мы извлекли из этого проекта, заключается в том, что понимание данных имеет решающее значение. Наборы данных могут быть ограничены по размеру, и они могут быть не изобразимы для всей совокупности, или в процессе сбора данных могут не учитываться возможные систематические ошибки. Систематические ошибки часто становятся очевидными только после тщательного анализа данных или при анализе связи между предсказаниями модели и входными данными модели. Понимание данных также определяет конструирование и выбор признаков, а также выбор алгоритмов и показателей производительности.
49. Модели контролируемого обучения также требуют хорошего уровня экспертных знаний для необходимой организации.
50. Мы осознали необходимость тестирования различных алгоритмов, выборок, например для разных периодов, характеристик/поведения и размеров. Все это может оказать существенное влияние на результаты.
51. Мы также понимаем, что нам может понадобиться ряд моделей, а не одна модель, для обучения и прогнозирования для различных подгрупп. Например, нам могут потребоваться отдельные модели для прогнозирования статуса ДММ для различных категорий мигрантов, например мигрантов, ранее приезжавших в страну, и мигрантов, приехавших впервые.

V. Дальнейшие шаги

52. Использование контролируемого машинного обучения для классификации ДММ по административным данным имеет некоторые уникальные потенциальные преимущества по сравнению с другими методами классификации, но также создает некоторые проблемы, к которым необходимо подходить с осторожностью.

- Построение признаков, чтобы максимально использовать информацию из данных проверок при выезде.
- Подробный анализ эффективности трех алгоритмов для когорт иммигрантов, включая исследование точности и смещения прогнозов между когортами и внутри них.
- Рассмотрение расширений контролируемого машинного обучения, которые включают аспекты анализа выживаемости.
- Оценка необходимости переобучения моделей с течением времени для устранения сдвига моделей, а также потенциальные подходы к мониторингу и переобучению.
- Разработка концептуальной модели для эмиграции.
- Изучение способов интеграции прогнозов ДММ с помощью машинного обучения в более широкую методологию НСС в области оценок миграции на основе административных данных (ОМАД) и как неопределенность этих прогнозов может распространяться и сообщаться пользователям.
- Применение аналогичных методов для классификации ДММ из стран ЕС в данных Базы данных регистрации и взаимодействий населения (RAPID)

53. Мы также хотели бы поработать по некоторым направлениям, которые в настоящее время недостаточно изучены, с использованием других методов машинного обучения. Например, алгоритмы неконтролируемого машинного обучения могут выявлять закономерности в административных данных и помогать формировать эвристический набор правил для круговой миграции и британских ДММ, где невозможно применить точно определенные метки ДММ. Проект пока что находится на ранней стадии, и мы приветствуем комментарии и предложения, которые помогут в этой работе.

VI. Список литературы:

54. Analysis Function (2021) Government Functional Standard GovS 010: Analysis, available from: https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/1011798/CO_Govt_Functional_Std_GovS010_Analysis_v2_Final_WEB.pdf
55. Government Statistical Services (2022) Reproducible Analytical Pipelines
56. Available from: https://gss.civilservice.gov.uk/reproducible-analytical-pipelines/
57. Home Office (2020a) Home Office statistics on exit checks: user guide, available from: https://www.gov.uk/government/publications/home-office-statistics-on-exit-checks-user-guide/home-office-statistics-on-exit-checks-user-guide
58. Home Office (2020b) Impact Assessment for changes to the Immigration Rules for Skilled Workers, available from: https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/936121/Revised_Impact_Assessment_for_the_Skilled_Worker_Route_signed.pdf
59. Office for National Statistics (2022a) Long-term international migration, provisional, year ending June 2021, available from https://www.ons.gov.uk/peoplepopulationandcommunity/populationandmigration/inte

rnationalmigration/bulletins/longterminternationalmigrationprovisional/june2021#main-points

60. Office for National Statistics (2022b) Estimates of overseas residents' visits and spending in the UK, available from https://www.ons.gov.uk/peoplepopulationandcommunity/leisureandtourism/datasets/estimatesofoverseasresidentsvisitsandspendingintheuk
61. Office for National Statistics (2021) Using statistical modelling to estimate UK international migration, available from https://www.ons.gov.uk/methodology/methodologicalpublications/generalmethodology/onsworkingpaperseries/usingstatisticalmodellingtoestimateukinternationalmigration
62. Office for National Statistics (2020) Exploring international migration concepts and definitions with Home Office administrative data, available from:
63. https://www.ons.gov.uk/peoplepopulationandcommunity/populationandmigration/internationalmigration/articles/exploringinternationalmigrationconceptsanddefinitionswithhomeofficeadministrativedata/2020-02-14
64. Office for National Statistics (2018) Report on international migration data sources
65. Pedregosa, Fabian; Varoquaux, Gaël ; Gramfort, Alexandre; Michel, Vincent; Thirion, Bertrand; Grisel, Olivier; Blondel, Mathieu; Prettenhofer, Peter; Weiss, Ron; Dubourg, Vincent ; Vanderplas, Jake; Passos, Alexandre; Cournapeau, David; Brucher, Matthieu; Perrot, Matthieu; Duchesnay, Édouard (2011) Supervised Learning overview, in Scikit-learn: Machine Learning in Python, JMLR 12, pp. 2825-2830, 2011.; 12(85):2825−2830, 2011. Available from: https://ogrisel.github.io/scikit-learn.org/sklearn-tutorial/tutorial/astronomy/general_concepts.html
66. Sapon, Muhammad & Ismail, Khadijah & Suehazlyn, Zainudin & Ping, Chew & Malaysia, Nasional & Lumpur, Kuala. (2022). Diabetes Prediction with Supervised Learning Algorithms of Artificial Neural Network.
67. United Nations (1998) Recommendations on Statistics of International Migration, Revision 1, available from: http://data.un.org/Glossary.aspx?q=long-term%20migrant
68. Zhang, LC. (2020) On provision of UK neighbourhood population statistics beyond 2021. arXiv [Preprint]. Available from: https://arxiv.org/abs/2111.03100