**МИНИСТЕРСТВО НАУКИ И ВЫСШЕГО ОБРАЗОВАНИЯ РОССИЙСКОЙ ФЕДЕРАЦИИ**

**ФЕДЕРАЛЬНОЕ ГОСУДАРСТВЕННОЕ АВТОНОМНОЕ ОБРАЗОВАТЕЛЬНОЕ УЧРЕЖДЕНИЕ ВЫСШЕГО ОБРАЗОВАНИЯ «СЕВАСТОПОЛЬСКИЙ ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ»**

Институт информационных технологий и управления в технических системах кафедра «Информационные системы»

**Пояснительная записка**

к проекту

«Разработка алгоритма прогнозирования выбора абитуриентом направления подготовки в вузе по цифровым следам в социальной сети»

по дисциплине «Проектная деятельность»

Выполнила студентка

Группы ИС-18-1-о

Попова Арина

Наставник

Куркчи Ариф Эрнестович

Севастополь

2020

**АННОТАЦИЯ**

В пояснительной записке описывается проект, в котором проводилась разработка прогнозирования выбора абитуриентом направления подготовки в вузе по цифровым следам в социальной сети. Цель проекта – построение модели, которая будет предсказывать направление обучение абитуриента по его профилю в социальной сети Вконтакте.

В документе приводится выявление и анализ проблемы, подробное описание предложенного решения и пошаговое описание его реализации. Также документ собирает и обобщает все сведения, которыми обладает команда за период работы над проектом.

**СОДЕРЖАНИЕ**

[ПОСТАНОВКА ЗАДАЧИ 3](#_Toc74834508)

[ИСХОДНЫЕ ДАННЫЕ 4](#_Toc74834509)

[ПОДГОТОВКА ДАННЫХ 11](#_Toc74834510)

[ПРИМЕНЕННЫЕ МЕТОДЫ 13](#_Toc74834511)

[ЗАКЛЮЧЕНИЕ 16](#_Toc74834512)

# **1 ПОСТАНОВКА ЗАДАЧИ**

Необходимо применить технологии анализа данных для исследования страничек абитуриентов в социальной сети Вконтакте с помощь языка программирования Python. Также необходимо создать датасет с помощью библиотеки Pandas для применения технологий машинного обучения, выполнив следующие этапы работ:

1) Классифицировать названия факультетов (2500) по предметным областям (полуавтоматическая и ручная разметка), данные о факультете выгружаются из профиля ВК абитуриента.

2) Выгрузить цифровой след абитуриентов, поступивших в бакалавриат.

3) Классификатором из п.1 определить направление подготовки в бакалавриате, по предметным областям, на которые поступил абитуриент.

4) Классифицировать образовательные интересы абитуриента по подпискам на сообщества в ВК, классификатор предметных областей из п.1 (полуавтоматическая и ручная разметка сообществ.

5) Выявление взаимосвязи между выбранным направлением в бакалавриате и цифровым следом из ВК: подписки на тематические сообщества, другие личные данные из профиля (пол, возраст, и др. поля).

Также с помощью ранее сформированного датасета построить модель машинного обучения, предсказывающую направление поступления абитуриента, выполнив следующие этапы работы:

1) Разбить данные на обучающую и тестовую выборку.

2) Создать автоматизированный pipeline подготовки данных к обучению, включающий в себя заполнение пропусков и создания новых признаков.

3) Обучить различные модели машинного обучения.

4) Выбрать лучшую, подобрав наилучшую комбинацию гиперпараметров.

5) Исследовать результаты по метрикам.

# **2 ИСХОДНЫЕ ДАННЫЕ**

Реализация задачи осуществляется с использованием языка Python с платформой Jupyter Notebook, а также программы Excel для хранения данных. Перед осуществлением анализа данных требовалось эти данные собрать и аккумулировать в необходимом для анализа и дальнейшего применения машинного обучения виде.

На первом этапе производилась разметка названий факультетов (около 2500 наименований) по предметным областям, а также областям наук. Пример названий факультетов приведен на рисунке 1.

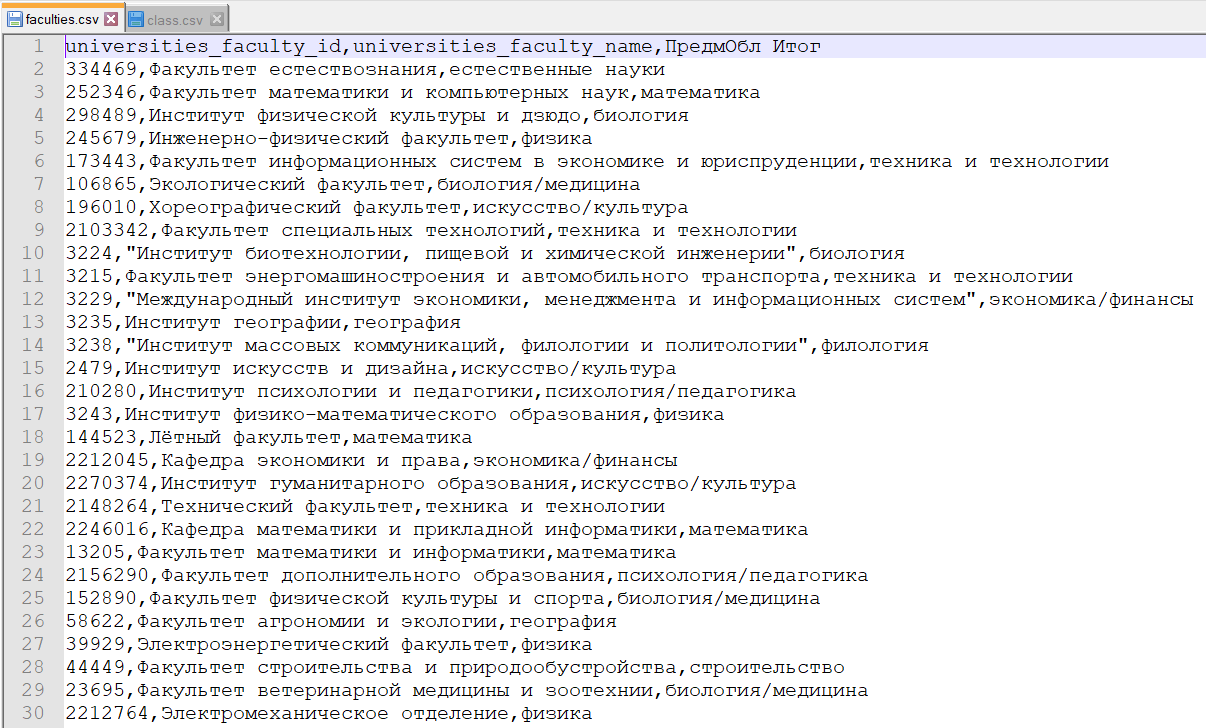


Рисунок 1 – Названия факультетов

На рисунке 2 приведен предложенный классификатор.

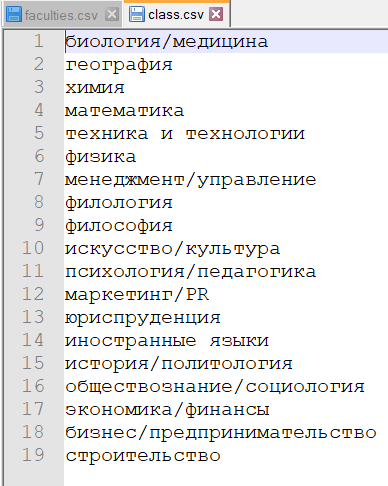


Рисунок 2 – Классификатор

Области же наук были разделены на следующие группы: гуманитарные, точные, естественные, а также другое.

В результате разметки получился файл, показанный на рисунке 3.

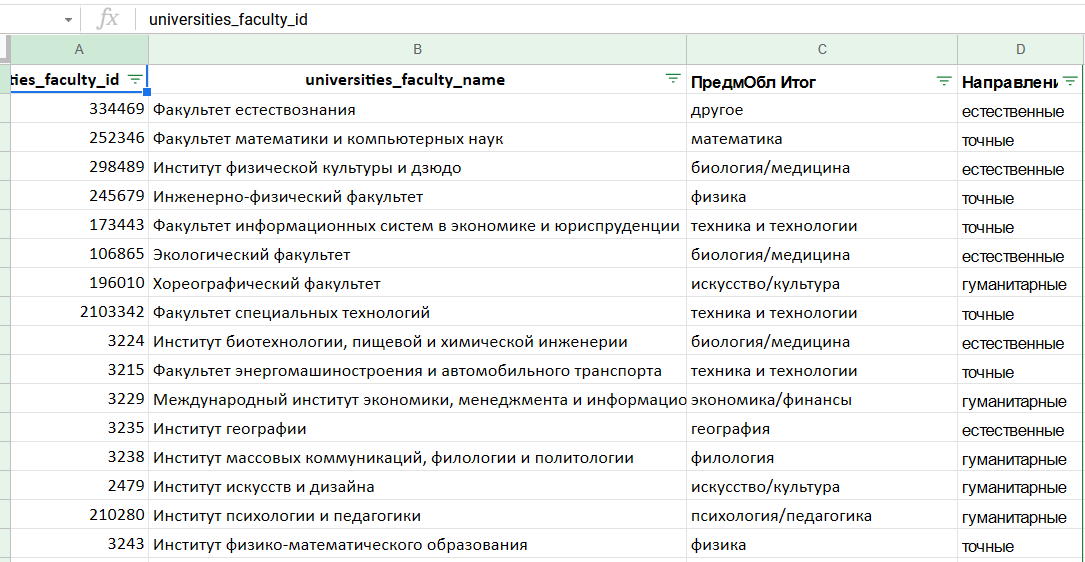


Рисунок 3 – Результат разметки

Следующим этапом работы являлось выгрузка необходимых данных с информацией о страничках пользователей в социальной сети Вконтакте, посту-пивших в бакалавриат, также определение их направления поступления, вы-грузка их подписок на сообщества, а также соединение таблиц для создания данных для исследования. Данные выгружались из базы университета автоматически, на выходе получался файл .csv. Таким образом, необходимо было обработать данные с id абитуриентов и их данным в социальной сети, таких файлов было несколько (2019 и 2020 года), затем идентифицировать по данным название факультета пользователя, затем, соединив таблички, определить к какому направлению относится пользователь с данным факультетом по построенному ранее классификатору. Исходные файлы показаны на рисунках 4-7.

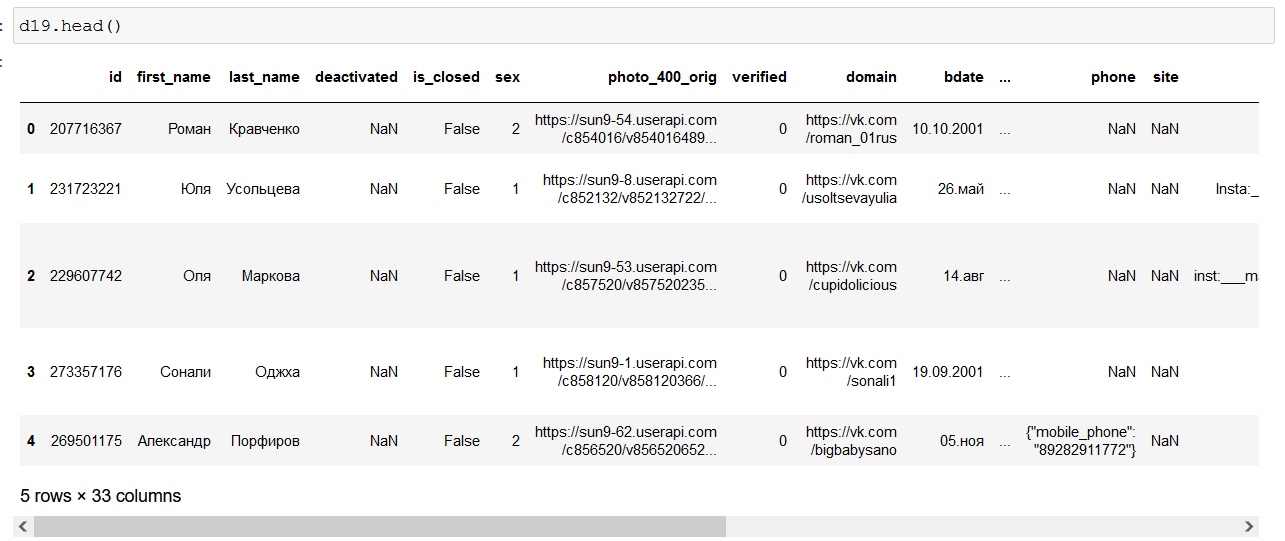


Рисунок 4 – Данные об абитуриентах за 2019 год

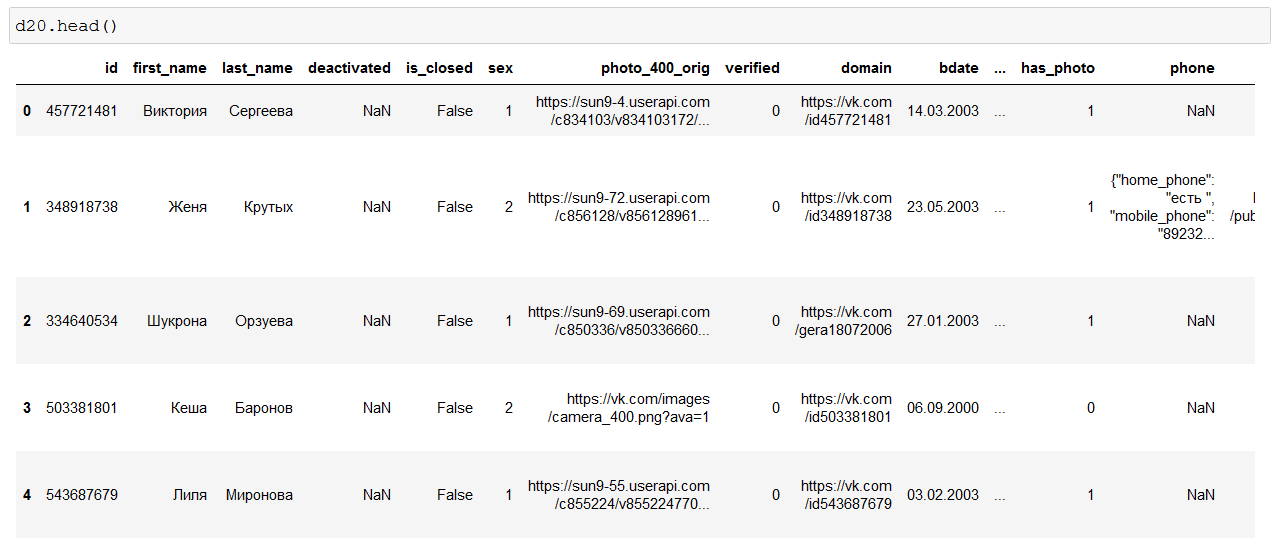


Рисунок 5 – Данные об абитуриентах за 2020 год

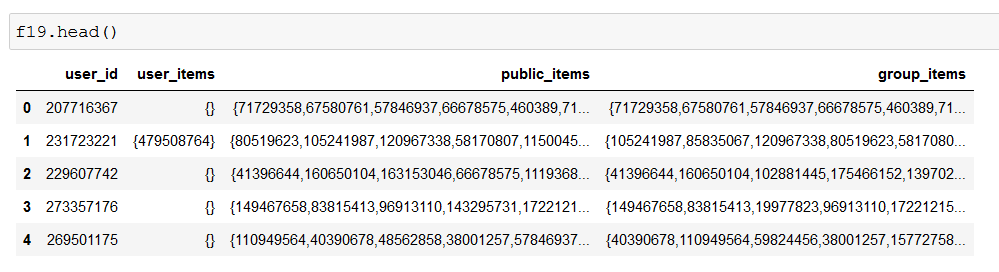


Рисунок 6 – Подписки абитуриентов 2019 года

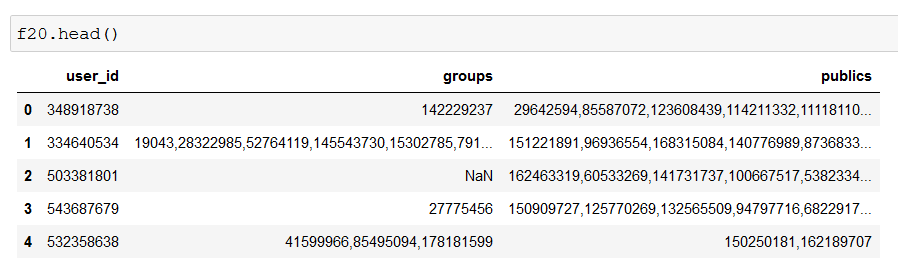


Рисунок 7 – Подписки абитуриентов 2020 года

В дальнейшем было произведено соединение таблиц абитуриентов разных годов, а также с таблицей, содержащей названия факультетов и пабликов. Результатом этой работы стало создание единого датасета.

Так как различных сообществ Вконтакте в данных было представлено свыше300000, то было принято решение использовать для анализа только те из них, что встречались хотя бы 100 раз, т. е. хотя бы 100 пользователей в данных были подписаны на данные сообщества (таких оказалось около 4 тысяч). Часть получившегося датасета показана на рисунке 8.

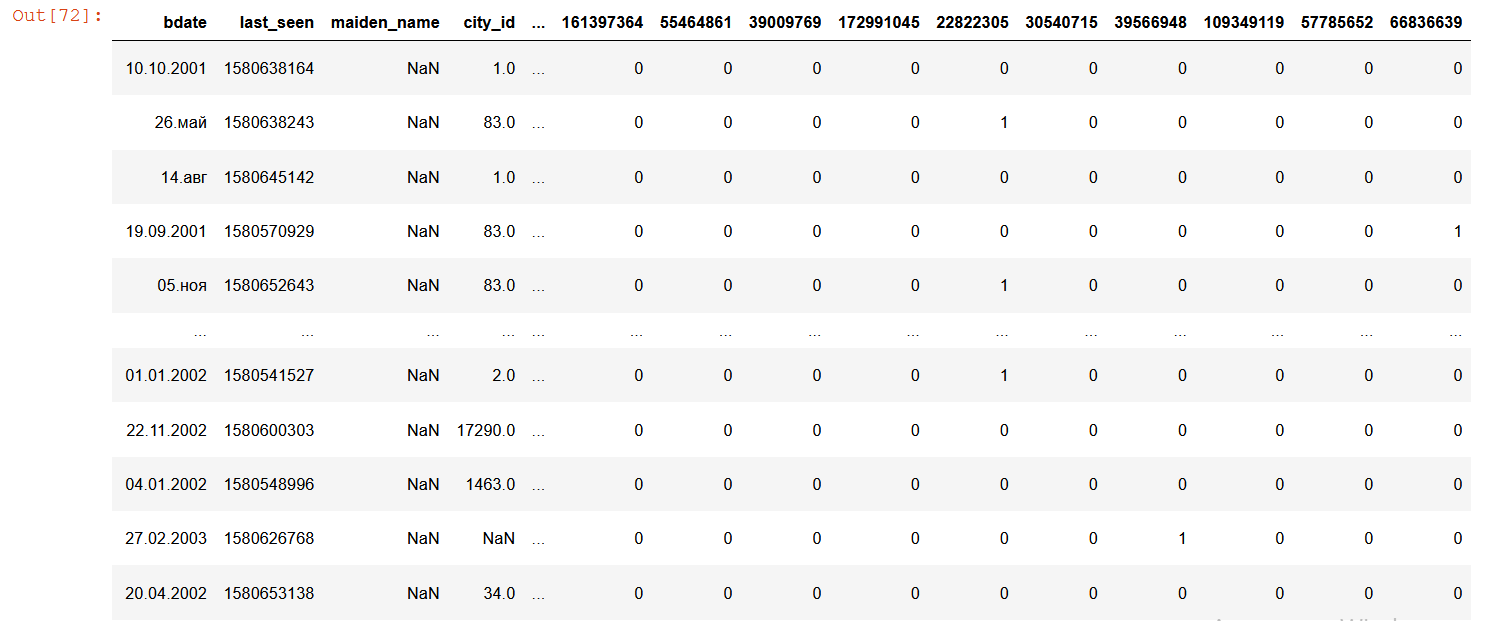


Рисунок 8 – Часть датасета

Последним и главным этапом работы оказалось исследование данных и со-здание новых признаков. Так, в процессе работы программы были созданы признаки об отношении абитуриента к алкоголю, курению, вероисповеданию, наличие девичьей фамилии и. т. д. Некоторые графики показаны на рисунках 9-13.

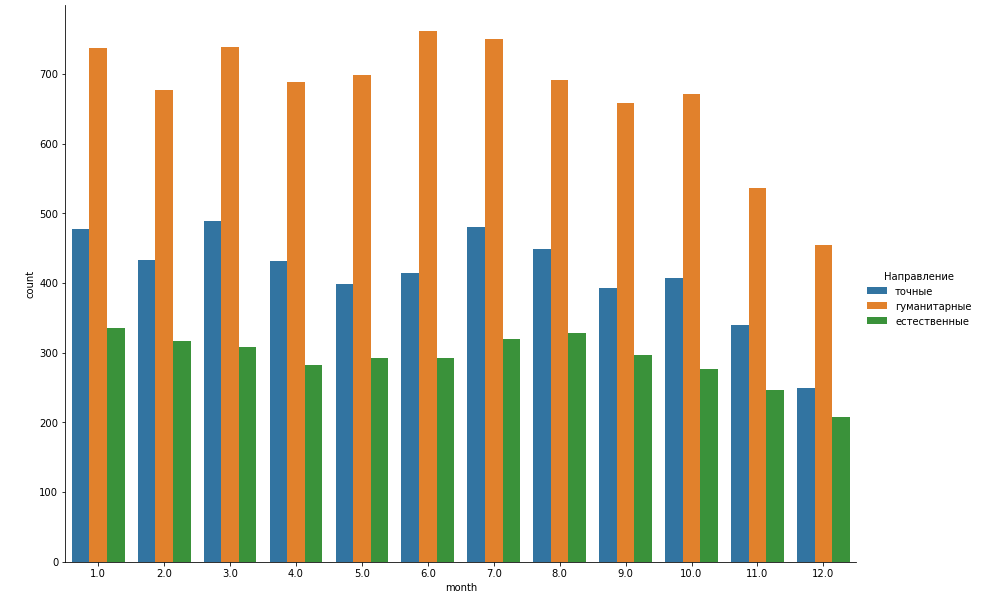


Рисунок 9 – График распределения направлений по номеру месяца рождения

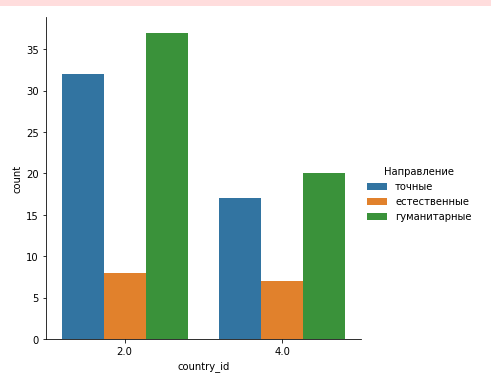


Рисунок 10 – График распределения направления по стране абитуриента

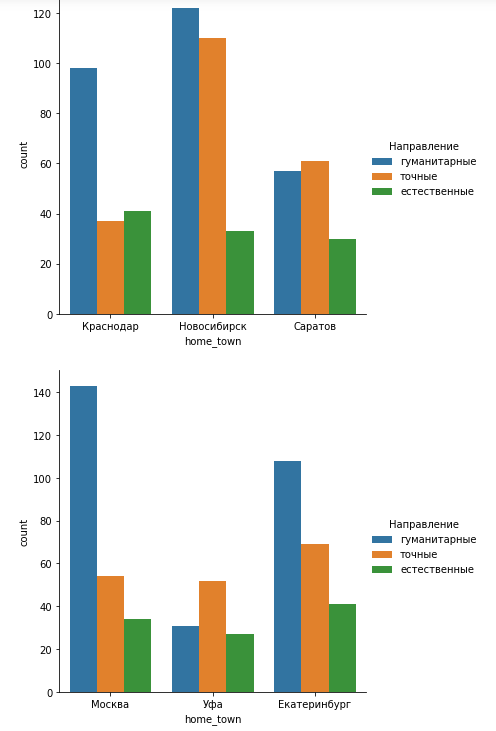


Рисунок 11 – График распределения направления по городу абитуриента

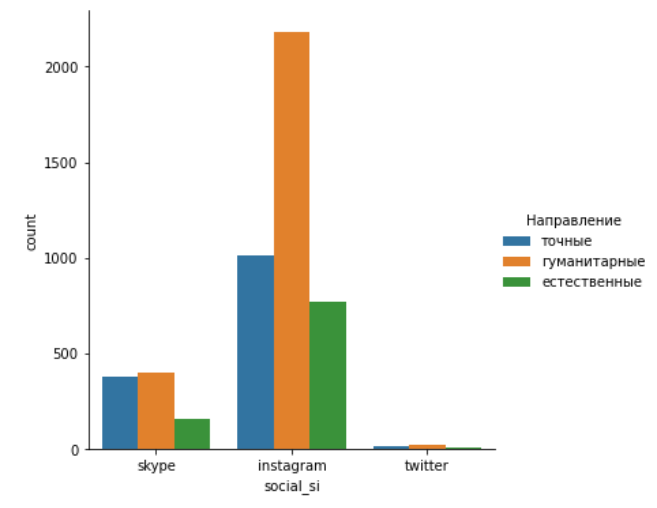


Рисунок 12 – График распределения направлений по указанным социальным сетям

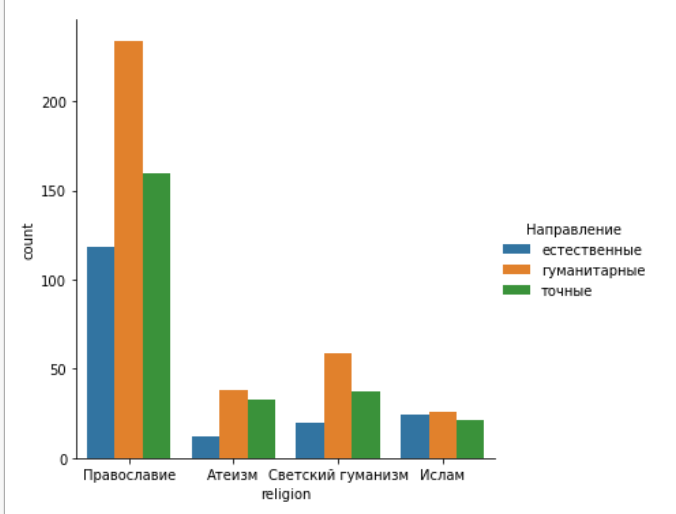


Рисунок 13 – График распределения направлений по религии абитуриента

После анализа признаков и создания новых был создан датасет, которые в дальнейшем будет использован для применения методов машинного обучения с целью построения предсказательной модели.

# **3 ПОДГОТОВКА ДАННЫХ**

Перед созданием модели машинного обучения необходимо было разделить выборку на обучающую и тестовую. Сделано это было с помощью вызова библиотеки sklearn и метода traintestsplit с параметром shuffle = True, т.е. и тестовая, и обучающая выборка стратифицированы по целевой переменной (гуманитарное, точное, естественное или другое направление). Разделение выборки на тестовую и обучающую показан на рисунке 14.

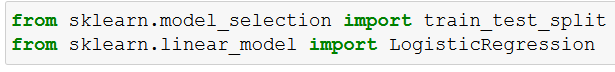


Рисунок 14 – Разделение выборки на тестовую и обучающую

Затем для удобства обучения были произведены манипуляции с данными с целью правильного расположения признаков друг за другом (сначала шли непосредственные данные человека по странице в Вконтакте, затем – преобразованные переменные и паблики.

В дальнейшем было необходимо настроить автоматическую обработку данных: заполнение пропусков в данных, создание категориальных переменных методом one-hot-encoding, причем отдельно должны обрабатываться категориальные и числовые переменные. Вызов библиотек для создания Pipeline показан на рисунке 15.

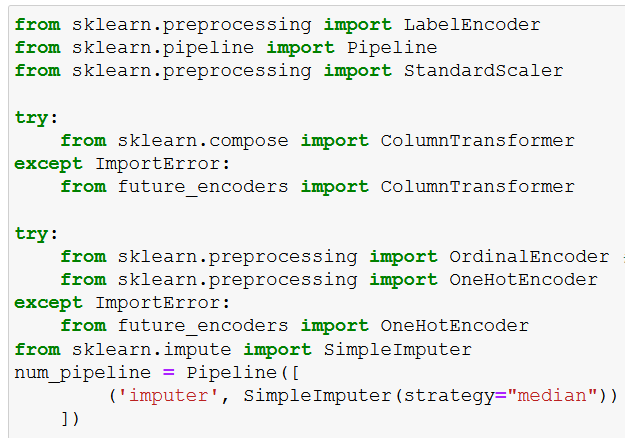


Рисунок 15 – Вызов библиотек для создания Pipeline

Отдельно была создана автоматическая обработка числовых признаков (numpipeline) и категориальных (catpipeline), затем все это объединялось в один fullpipeline. Категориальные и числовые признаки показаны на рисунке 16.

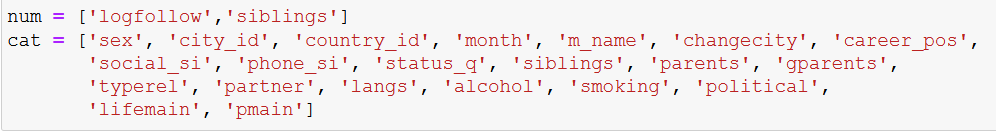


Рисунок 16 – Числовые и номинальные переменные

Как числовые, так и категориальные переменные в своих автоматических обработках заполняли пропуски. В категориальных переменных было использован самый популярный по столбцу, для номинальной – медианное значение, поскольку признаки не обладали незначительными отклонениями от нормальности. Важным отличием обработки категориальных переменных было создание новых столбцов переменных на основе номинальной переменной, для этого специально был написан класс-обработчик, показанный на рисунке 17.

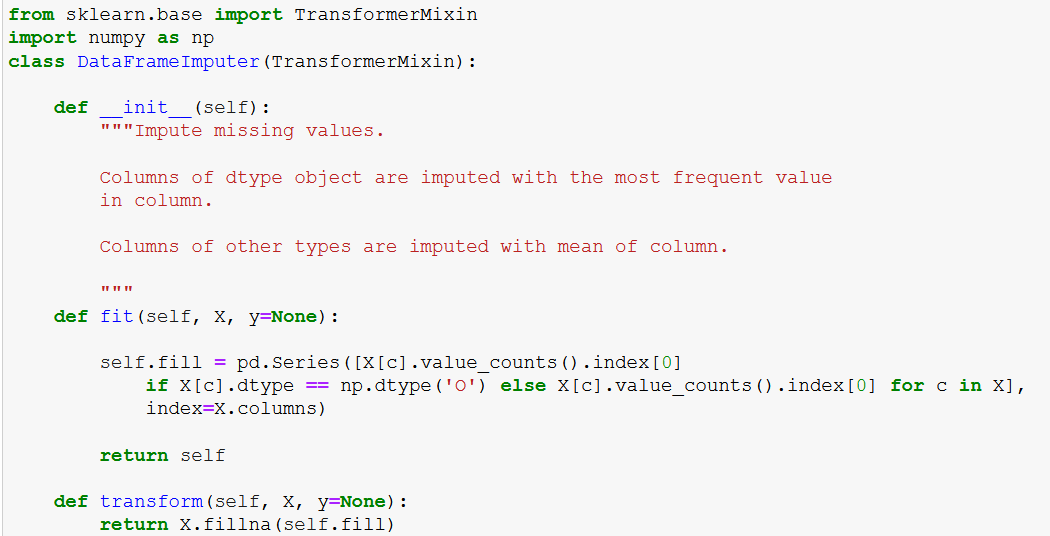


Рисунок 17 – Обработка категориальных признаков

Таким образом, с помощью pipeline как в тестовых, так и обучающих данных заполнялись пустые значения, а также создавали новые признаки на основе категориальных переменных.

# **4 ПРИМЕНЕННЫЕ МЕТОДЫ**

Затем строились различные модели, среди которых были LogisticRegression и RandomForest, лучшей оказалась последняя. Гиперпараметры были найдены с помощью метода кросс-валидации. Подбор гиперпараметров показан на рисунке 18.

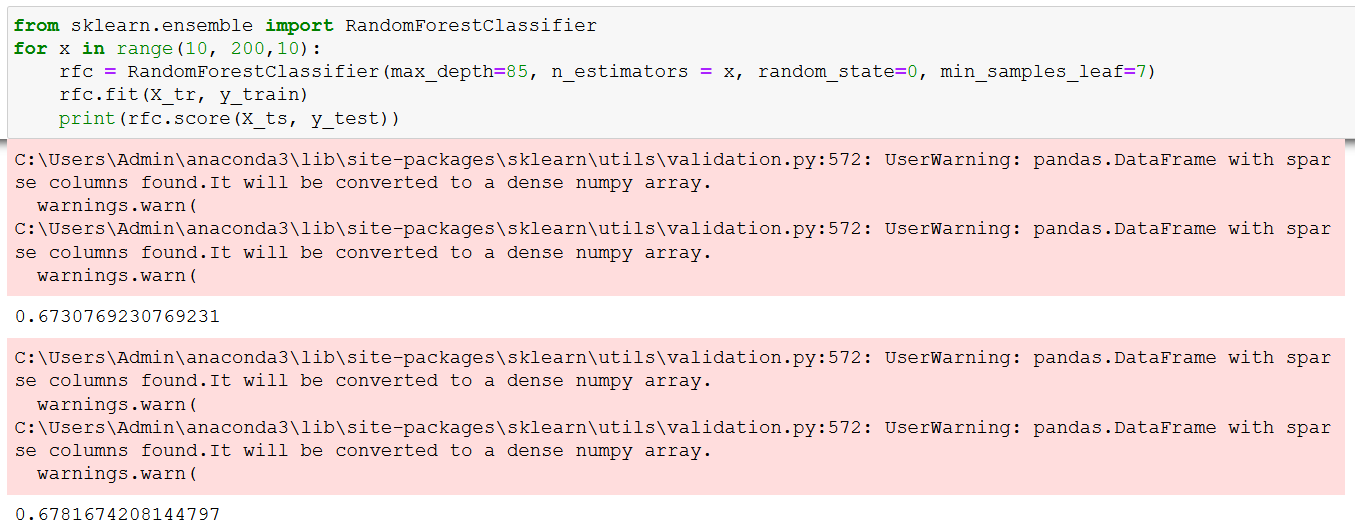


Рисунок 18 – Подбор гиперпараметров

Так как признаковое пространство обладает немалыми размерами, что делало затратной по времени валидацию, было принято решение перебирать параметры в широком диапазоне, а затем, найдя лучшую комбинацию гиперпараметров, с помощью перебора в цикле двух самых важных параметров –nestimators (количество деревьев) и maxdepth (максимальная глубина дерева).

Затем были проанализированы результаты различных метрик на обучающей и тестовой выборке. В ходе подробного рассмотрения различных метрик на тестовой и обучающей выборках, было выяснено, что гуманитарии распознаются моделью хорошо, в то время как класс другое совсем не распознается. Одна из метрик показана на рисунке 19.

Тогда удалось обучить слабую модель, которая отлично распознает класс «другое». Такой стала модель случайного леса со своим набором гиперпараметров. Метрика новой модели показана на рисунке 20.

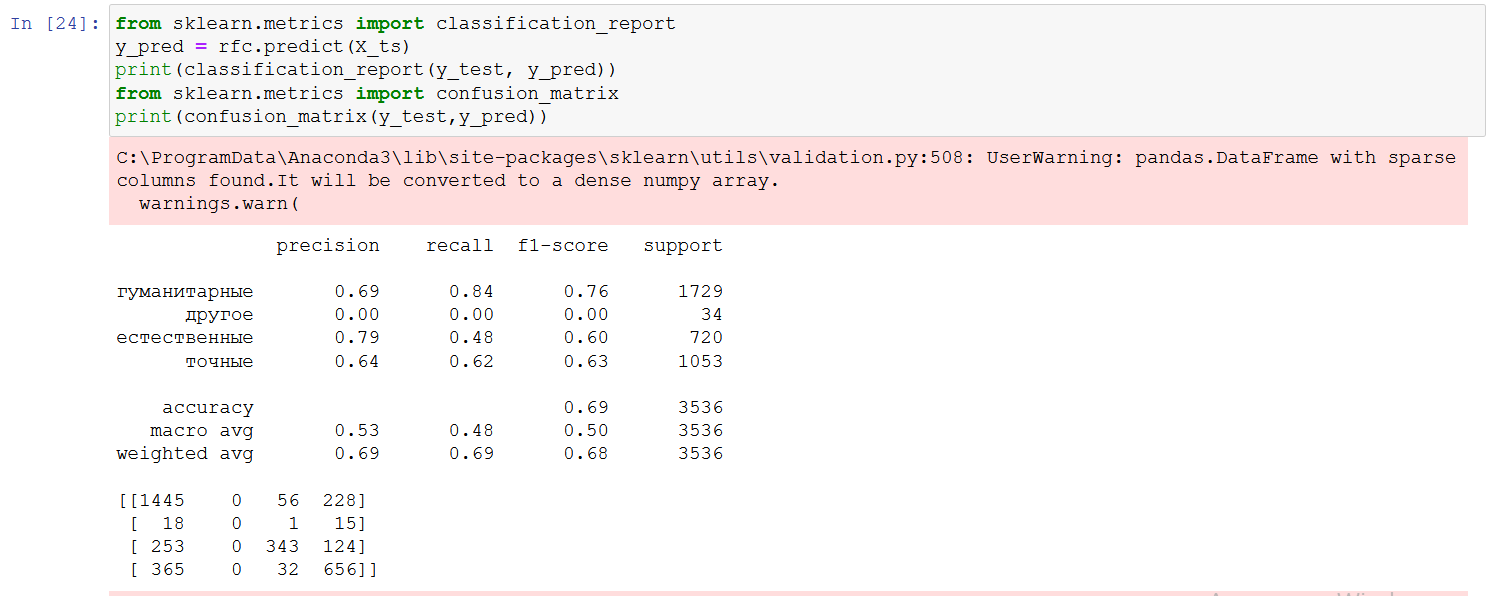


Рисунок 19 – Матрица ошибок

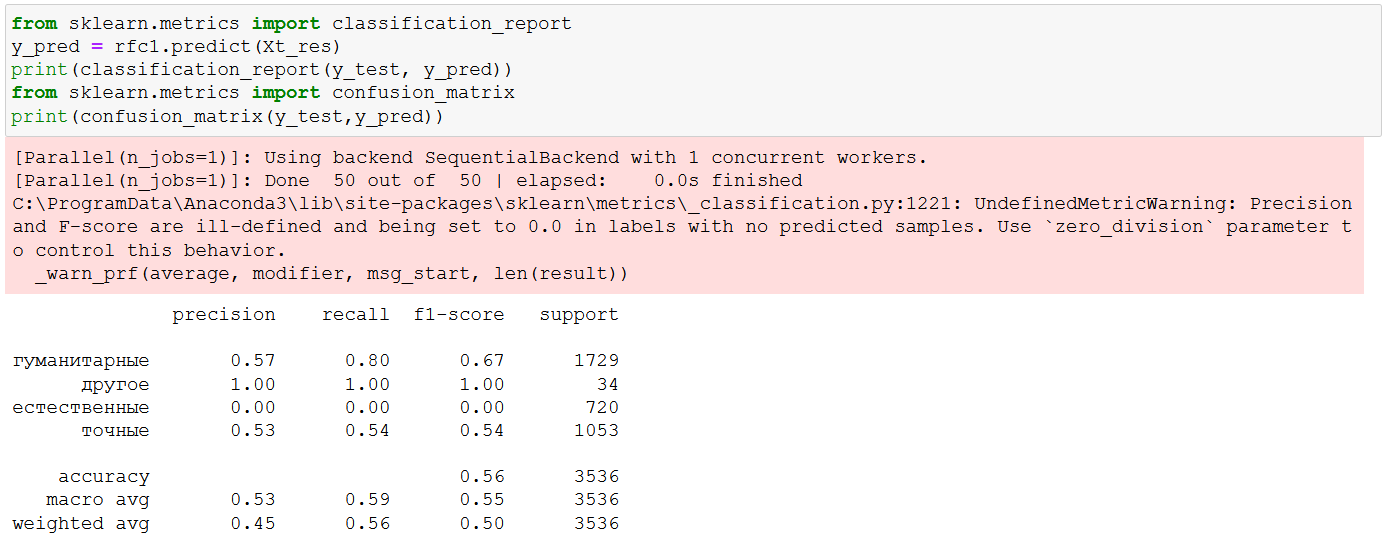


Рисунок 20 – Модель для класса «другое»

Затем, путем объединения двух лучших моделей, хорошо распознающих класс гуманитариев и другое, получили наилучшую модель с метрикой accuracy 0.70. Результаты этой метрики для итоговой модели показаны на рисунке 21.

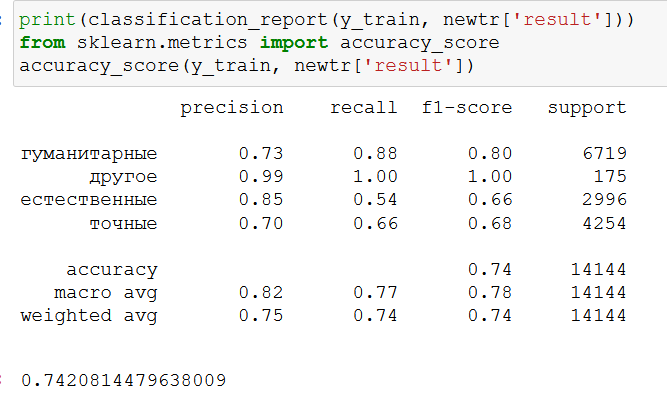


Рисунок 21 – Итоговая модель

# **ЗАКЛЮЧЕНИЕ**

Финальные результаты оценивались по метрике ROC-AUC, предназначенной для оценки моделей с несбалансированными классами представлены на рисунке 22.

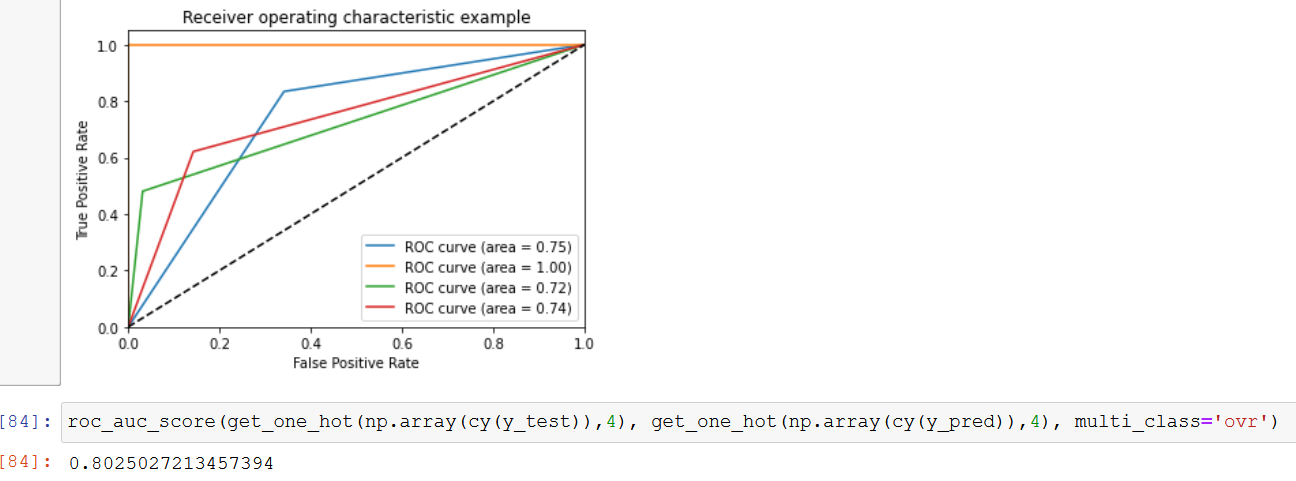


Рисунок 22 – Метрика ROC-AUC

Метрика ROC-AUC, равная 0.8 свидетельствует о хорошем качестве предсказаний, хотя достигнут он в большей степени из-за практически идеального предсказания класса другое, а также за счет хорошего распознавания класса гумманитариев, предсказания моделей насчет точников и естественников оставляют желать лучшего.