

Первичный анализ оттока клиентов с помощью Python

Романов Даниил Алексеевич

Приамурский государственный университет имени Шолом-Алейхема

Студент

Аннотация

Целью данной статьи является, анализ данных о клиентах банка и выявление на их основе результата. Программа будет написана в бесплатной интерактивной облачной среде для работы с кодом google colab с использованием языка программирования Python и его библиотек pandas, numpy, plotly, seaborn, sklearn. Результатом исследования станет вывод из анализа - какие клиенты, какого возраста и пола наиболее подвержены оттоку.

Ключевые слова: анализ данных, отток клиентов, Python, google colabe, plotly, pandas, numpy, seaborn, sklearn

Initial analysis of customer churn using Python

Romanov Daniil Alekseevich

Sholom-Aleichem Priamursky State University

Student

Abstract

The purpose of this article is to analyze data about bank customers and identify results based on them. The program will be written in a free interactive cloud environment to work with google colab code using the Python programming language and its pandas, numpy, plotly, seaborn, sklearn libraries. The result of the study will be the conclusion from the analysis - which clients, what age and gender are most susceptible to outflow.

Keywords: data analysis, customer churn, python, google colabe, plotly, pandas, numpy, seaborn, sklearn

1 Введение

1.1 Актуальность

Python - один из наиболее часто используемых языков программирования для анализа финансовых данных, с множеством полезных библиотек и встроенной функциональностью. В данной статье будет показано, как библиотеки машинного обучения Python можно использовать для прогнозирования оттока клиентов. Отток клиентов — это финансовый термин, который относится к потере клиента, то есть когда клиент перестает взаимодействовать с компанией или бизнесом. Аналогичным образом, скорость оттока — это скорость, с которой клиенты покидают компанию в

течение определенного периода времени. Уровень оттока, превышающий определенный порог, может иметь как материальные, так и нематериальные последствия для успеха бизнеса компании. В идеале, компании хотели бы удержать как можно больше клиентов. С появлением передовых технологий обработки данных и машинного обучения компании теперь могут выявлять потенциальных клиентов, которые могут прекратить вести с ними бизнес в ближайшем будущем. В данной статье будет рассмотрен примеер, как с помощью библиотек Python можно сделать первичный анализ данных.

1.2 Обзор исследований

В своей работе В.С. Мхитарян, М.Ю. Ахипова и Т.А. Дуброва описывают теоретические и алгоритмические основы анализа данных [1]. В.А. Мельникова, Д.А. Медведев рассказывают почему именно Python чаще всего используется в анализе данных и приводят этому ряд примеров и аргументов [2]. А.А. Карякина, А.В. Мельников сравнивают модели прогнозирования для оттока клиентов интернет провайдеров [3].

1.3 Цель исследования

Цель исследования – с помощью языка программирования Python и его библиотек для анализа данных проанализировать таблицу данных и выявить какие клиенты наиболее подвержены оттоку.

2 Материалы и методы

Для написания программы используется среда программирования google colab [4] и таблица с данными [5].

3 Результаты и обсуждение

Перед началом работы следует зайти в свой аккаунт google colab и загрузить в него таблицу с данными. После чего импортируем нужные библиотеки (рис.1).

```
import plotly
import plotly.graph_objs as go
import plotly.express as px
from plotly.subplots import make_subplots

import numpy as np
import pandas as pd
import plotly.io as pio

from sklearn.model_selection import train_test_split
from sklearn import svm
import seaborn as sns
import matplotlib.pyplot as plt
from sklearn.metrics import mean_squared_error
from sklearn.naive_bayes import GaussianNB
```

Рисунок 1 - Импортирование нужных библиотек

Загружаем набор данных в переменную с помощью pandas и открываем его для предварительного просмотра (рис.2).

```
# Загружаем данные с помощью pandas
df_churn = pd.read_csv(open("Churn_Modelling (1).csv","rb"))
df_churn.head()
```

Рисунок 2 - Загрузка данных

С помощью plotly отобразим график, показывающий сколько активных клиентов мужского и женского пол. Из графика видно, что среди активных пользователей немного больше мужчин (рис.3).

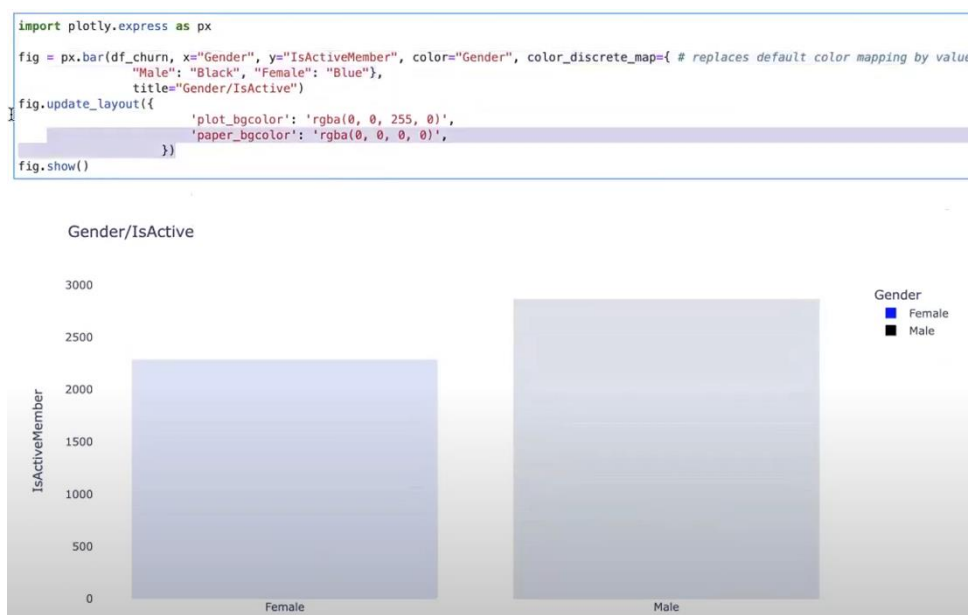


Рисунок 3 - График распределения по гендеру

Отообразим график показывающий активных клиентов по возрасту. По графику видно, что наибольшее число активных клиентов находятся в возрасте от 30 до 40 лет (рис.4).

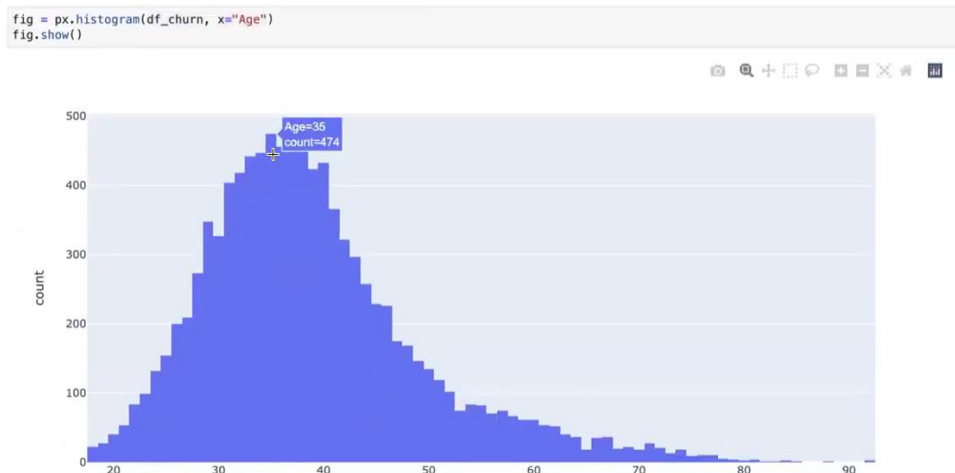


Рисунок 4 - График возраста

Построим график по географическому местоположению (рис.5).

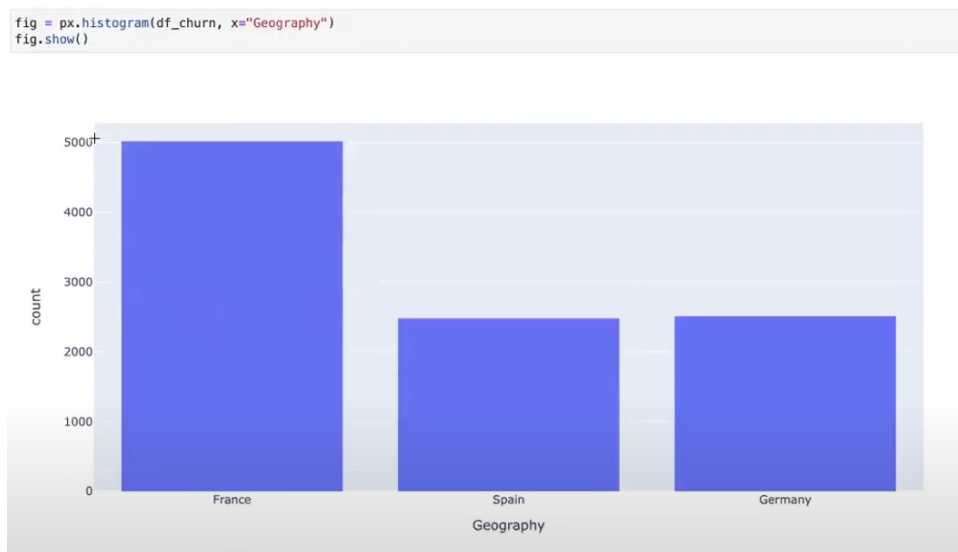


Рисунок 5 - График географического распределения

Создаём график по кредитной истории (рис.6).

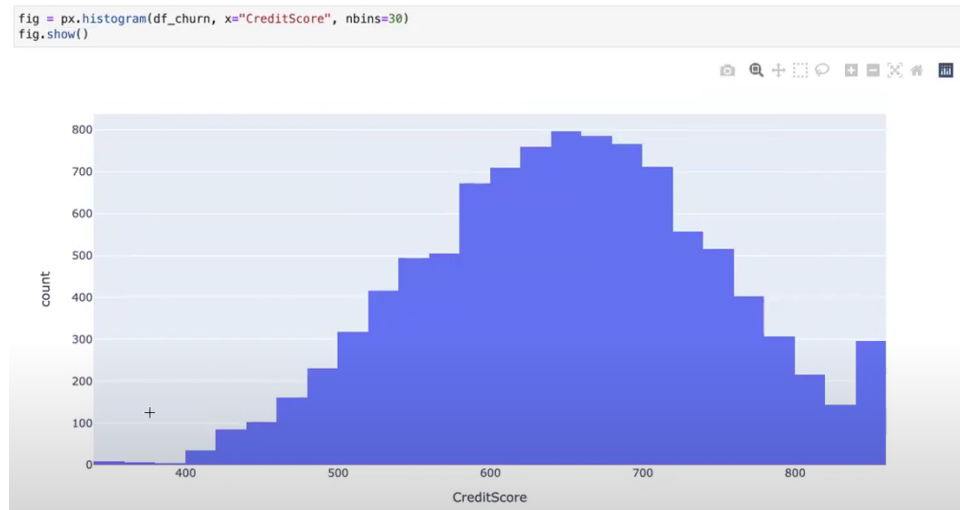


Рисунок 6 - График кредитной истории

С помощью метода `scatter` можно построить гистограмму двумерного распределения, показывающую взаимосвязь возраста клиентов и их кредитных историй (рис.7).

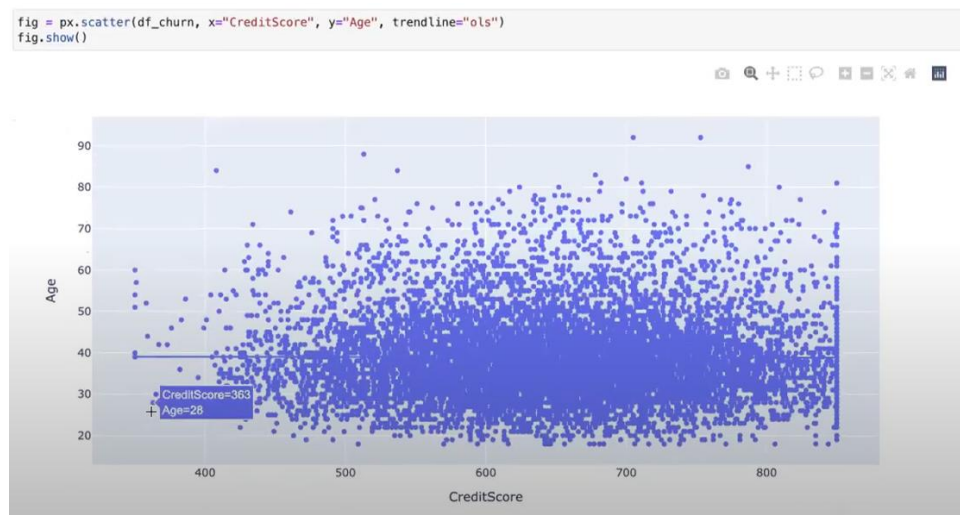


Рисунок 7 - Гистограмма распределения возраста и кредитной истории

Тоже самое можно сделать, используя характеристики - баланс на счёте и возраст (рис.8).



Рисунок 8 - Гистограмма возраста и баланса на счёте

Переходим к завершающей гистограмме, которая отобразит количество активных и неактивных клиентов по полу и возрасту. Атрибут метода `scatter` – `facet_col` создаёт 2 графика по заданным признакам. Исходя из графика видно, что большая часть оттекающих клиентов находятся в возрасте от 45 лет, как мужчины, так и женщины (рис.9).



Рисунок 9 - Гистограмма зависимости оттока от возраста и пола

Выводы

Благодаря Python и сопутствующим библиотекам по анализу данных удалось проанализировать данные клиентов и выяснить, что большая часть оттекающих клиентов как мужчины, так и женщины, находятся в возрасте от 45 лет.

Библиографический список

1. Мхитарян В. С. и др. Анализ данных. Учебник. М.: Издательство Юрайт,

2018. 490 с..

2. Мельникова В. А., Медведев Д. А. Анализ больших данных с использованием Python // Труды Братского государственного университета. Серия: Естественные и инженерные науки. 2019. Т. 1. С. 46-49.
3. Карякина А. А., Мельников А. В. Сравнение моделей прогнозирования оттока клиентов интернет-провайдеров // Машинное обучение и анализ данных. 2017. Т. 3. №. 4. С. 250-256.
4. Google colab URL: <https://colab.research.google.com>
5. Таблица с данными URL: <https://disk.yandex.ru/d/VSF91KroML2aSA>