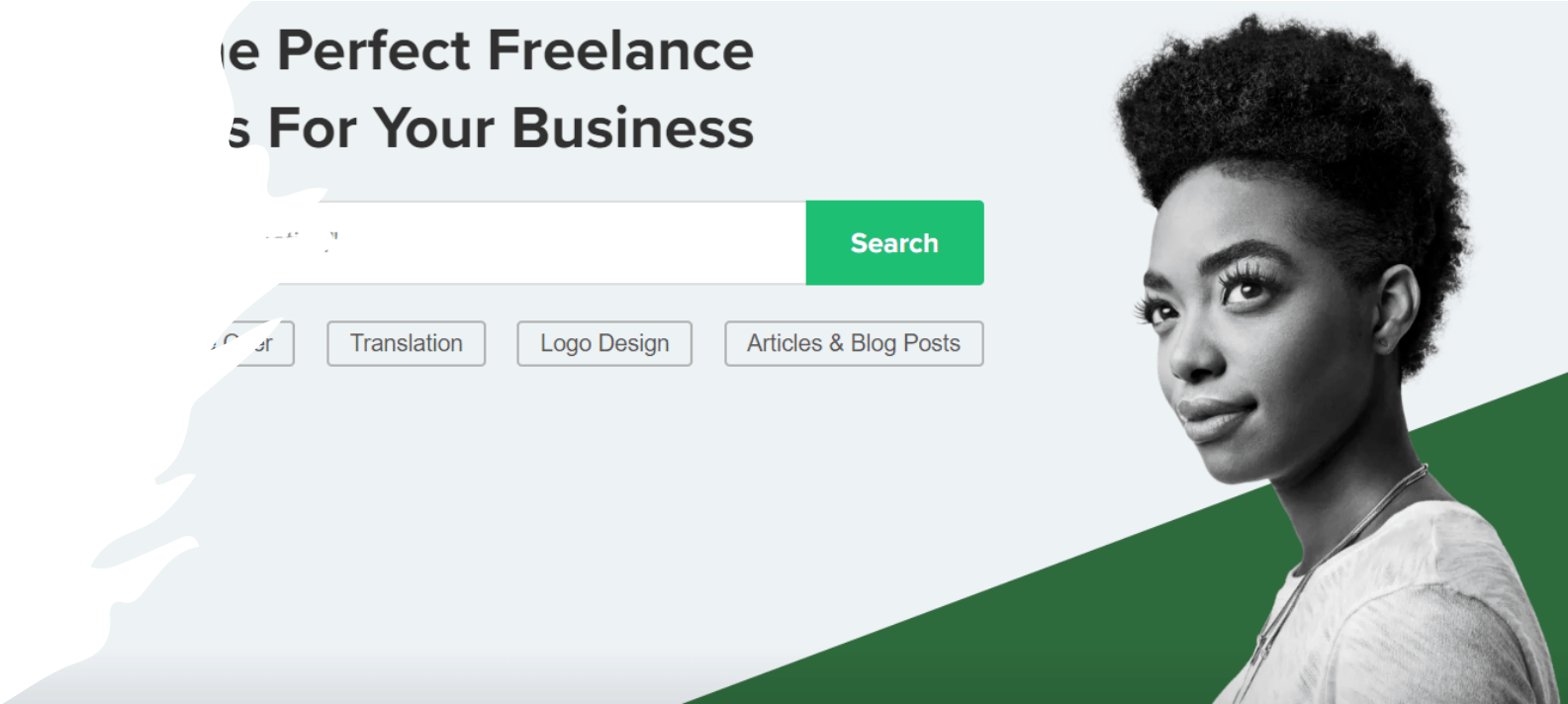
פרוייקט גמר במבוא למדעי הנתונים

מגישים: ליאב מורדוך, אדם לנצמן

נושא העבודה: Fiverr



הגדרת הבעיה ושאלת המחקר

כשאנחנו פותחים גיג בפייבר, קשה מאוד לדעת מראש האם נצליח או ניכשל, ואם כן, כמה או מה נוכל לשפר כדי להגדיל את הסיכויים שלנו להצליח, לכן אנחנו שאלנו את עצמנו את שאלת המחקר הבאה:

האם ניתן לחזות האם גיג יהיה מוצלח ואת מידת ההצלחה (ציון משוקלל של כמות הביקורות והדירוג) של גיג על פי נתוני הגיג (קטגורייה, מחירים, יכולות, שפה, רמת המוכר וכד..)?

```
~ python -c "import requests; from pprint import pprint; pprint(requests.get('https://www.fiverr.com/').text)" | grep "human"
'content="noindex"><title>It needs a human '
'data-identifier="title">It needs a human touch</h1><article><p '
'human touch","content":"Complete the task and we\'ll get you right back into '
```

זיהוי הנתונים והרכשתם

ב2023, לפייבר היו בערך 5 מיליון קניות, ולפי בדיקה שלנו בהחלט יש כמות דומה לזה של הצעות servicesו שאנשים מציעים. לנו אין צורך לקחת את כולם. אבל אנחנו עדיין רצינו להיות מגוונים בלקיחה, ולכן לקחנו מדגם של gigs מכל קטגוריה (המדגם מסתכם ב122808 נתונים לפני ניקוי, נתון לדיבייט שאף יותר מכיוון שחלק מהנתונים הם מערכים).

לפייבר יש הגנה מאוד מסובכת שבאה למנוע crawling. לכן כל בקשה שנעשה עם requests/aiohttp לא באה בחשבון. (בתמונה ניתן לראות את הדבר אפילו על בקשה בסיסית לurl)

הפתרון שלנו להרכשת הנתונים

ישר חשבנו על selenium. אך למרבה הפתעתנו - גם selenium עם כל שילוב של user-headers שחשבנו עליו לא עבד והביא אותנו למסך של אימות ריבוט (שלא היה פתיר גם לבני אדם - אגב).

אחרי הרבה חיפוש, נתקלנו בספרייה selenium-stealth, שהצליחה להסתיר את עצמה מספיק כדי לאפשר לנו את החיפוש. אך הבעיה היא, שהספרייה לא עודכנה מעל 3 שנים - ולכן היה לה בעיות יציבות. אחרי בערך 3-4 בקשות, הדרייבר הפסיק לפעול. היה לנו הרבה גישות לפתור את הבעיה, בין אם לבדוק את מצב הדרייבר כל כמה זמן (לא אפשרי - הדרייבר תקוע ולא מגיב, תוקע את כל התוכנה), לבין אם להריץ את הפונקציה כמה פעמים. ובסוף הפתרון ה"פשוט" ביותר היה האפקטיבי ביותר

```python
def timeout_callback(category_url):
    os.system("taskkill /F /IM chrome.exe /T")
    sys.exit()
```

```python

def crawl_gig(gig_url):
    timer = threading.Timer(18 , timeout_callback, [gig_url])
    timer.start()
    try:
```

```python

            category = soup.select("span.category-breadcrumbs")[-1].select_one("a").text
        except:
            timer.cancel()
            return None
        timer.cancel()
    except Exception as e:
        sys.exit()
    return {
```

```
@echo off

set "gigs_file=gigs.txt"
set "python_script=get_gigs.py"
set "timeout_seconds=5"

:loop
set /p first_line=<"%gigs_file%"
timeout /t %timeout_seconds% /nobreak >nul
python "%python_script%"

set /p new_first_line=<"%gigs_file%"
if "%new_first_line%"=="%first_line%" (
    call :removeDuplicateLine
)

goto loop

:removeDuplicateLine
ren "%gigs_file%" "gigs_temp.txt"
for /f "skip=1 delims=" %%a in (gigs_temp.txt) do echo %%a>>"%gigs_file%"
del "gigs_temp.txt"
exit /b
```

ניתוח ראשוני וטיוב נתונים

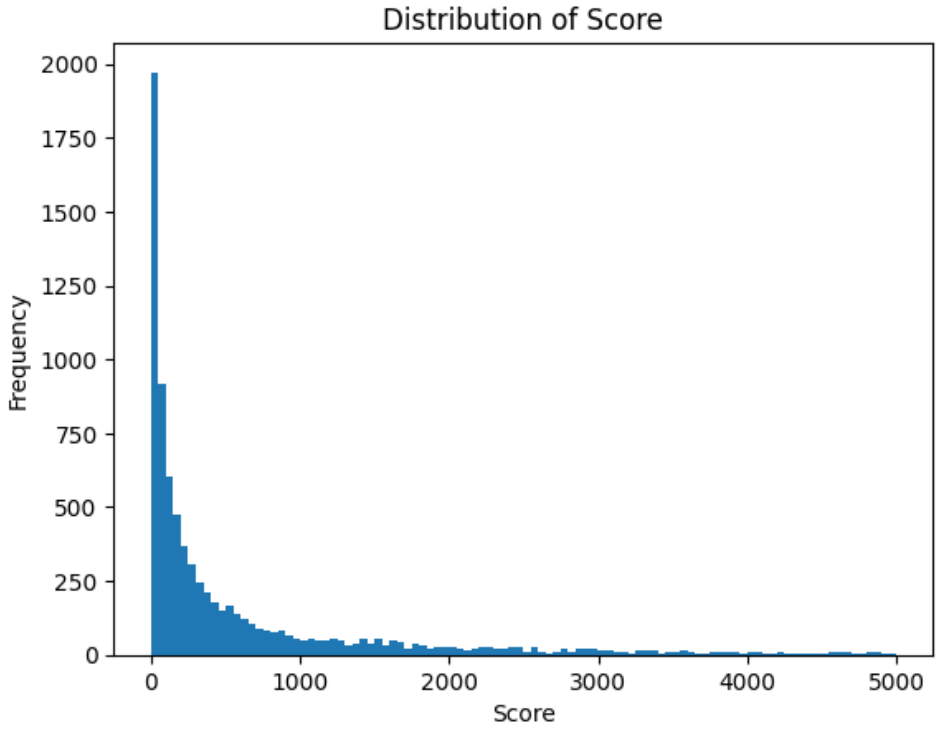
אחרי מחיקת כפיליות (בערך 500~ שורות - עדיין נשארנו עם K116).

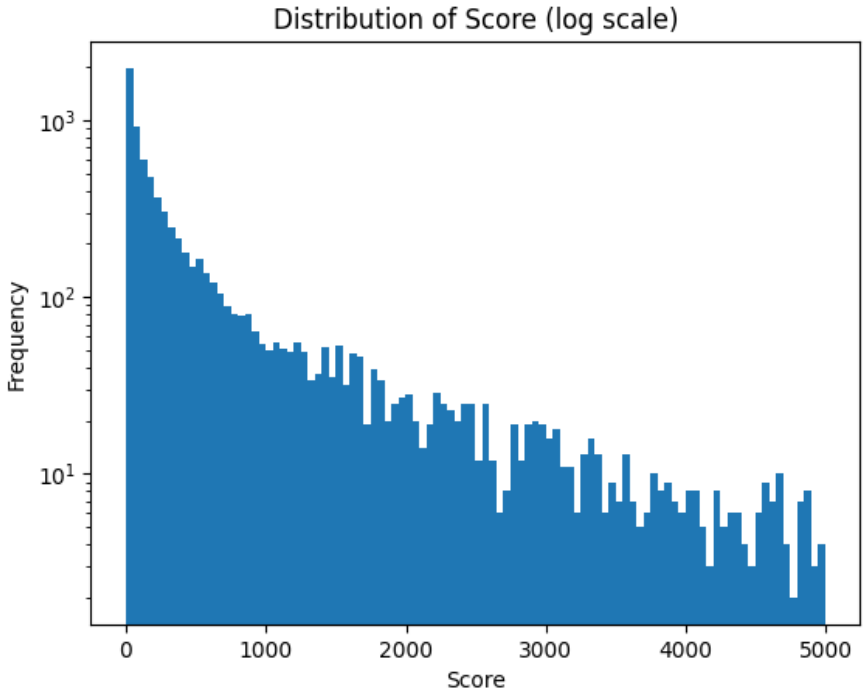
היה עלינו לפתור בעיה בתצוגה, בגלל שהמידע נלקח כנראה בצורה מוזרה (כי עברית ותכנות לא מסתדר), כל מיני סימנים השתנו (הופיעו "ªâ", כנ"ל לגבי החודשים "éåð" במקום יוני, "÷àå" במקום אוקטובר..

היה עלינו גם לתרגם מילים כמו unlimited לinfinity לתרגם את הרמה של המוכרים ועוד (פרטים מלאים בnotebook).

EDA ו-וויזואליזציה

בגלל שלא היה לנו דרך להגדיר האם gig הוא טוב או רע וכמה, היינו צריכים ליצור מערכת scoring משל עצמנו. המערכת היא חישוב קל של הדירוג (1-5 כוכבים) כפול כמות הreviews. ככה שההצלחה לוקחת בחשבון די הכל וגם - פשוטה להבנה.



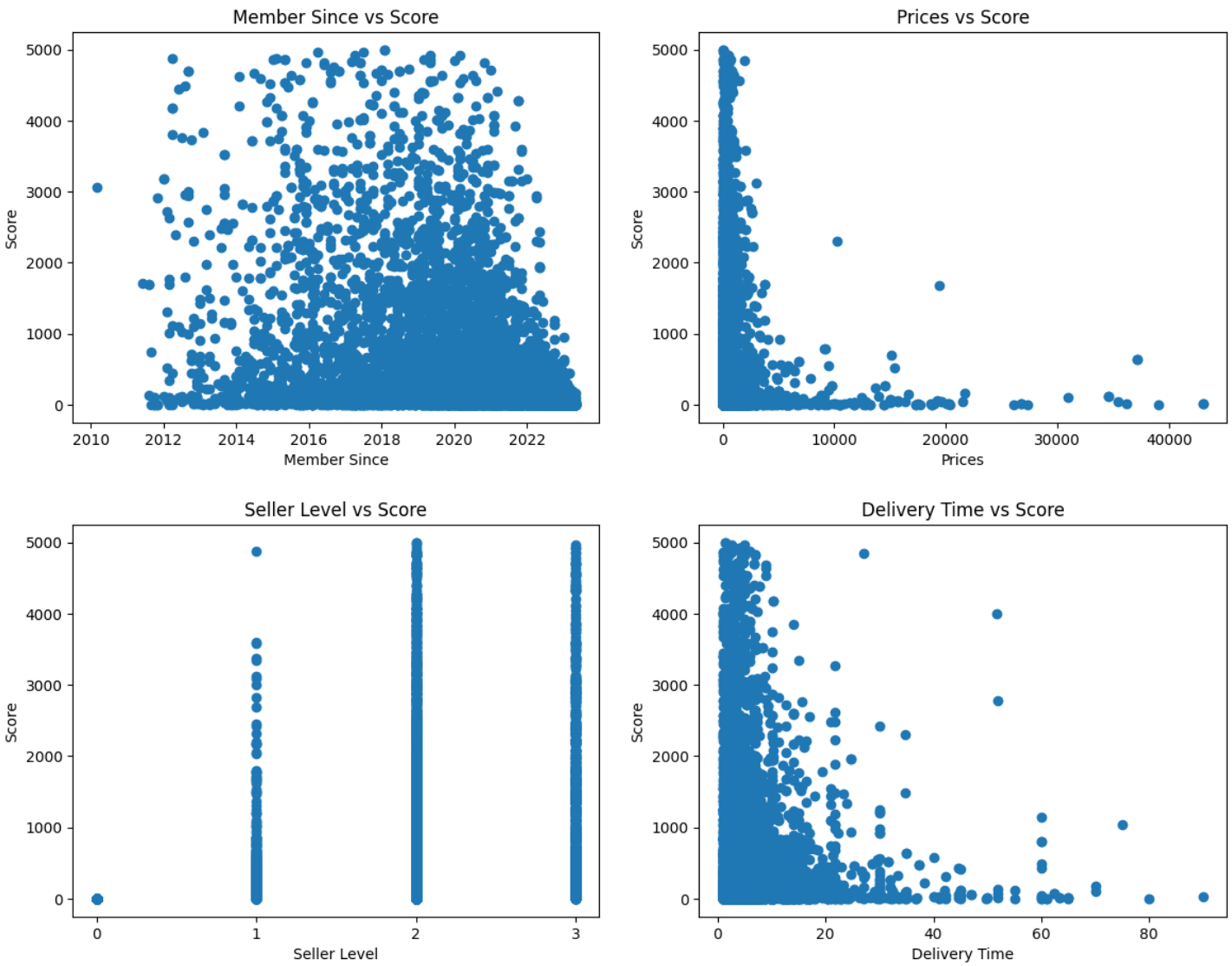


20 Most Common Words in Gig Titles

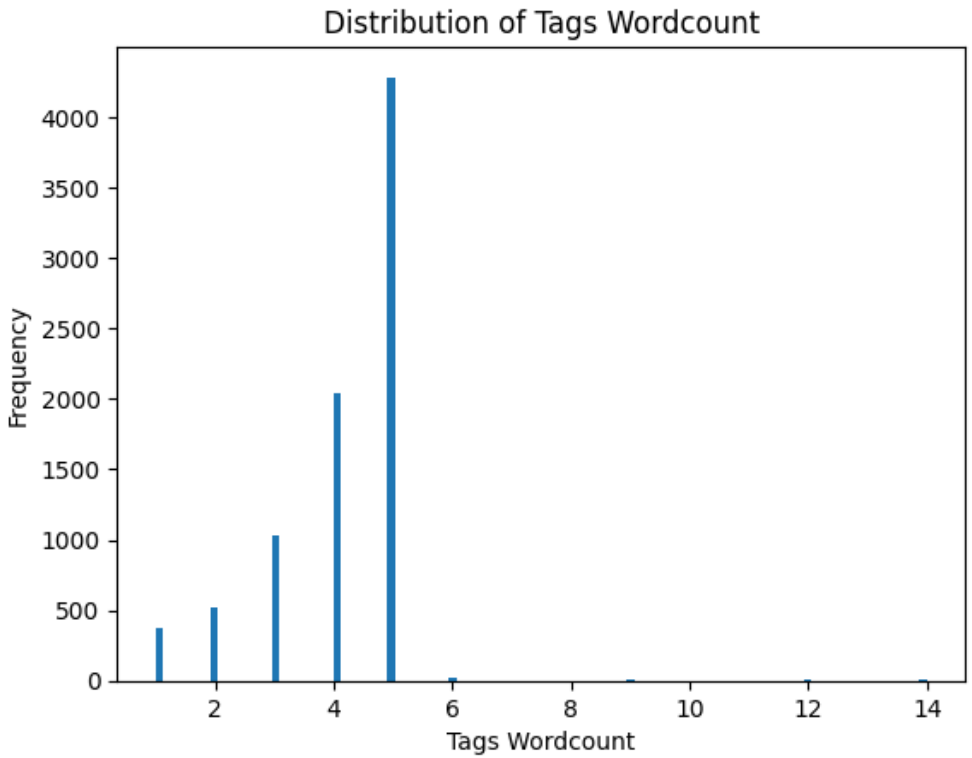
20 Most Common Words in the 100 Best Gigs by Score

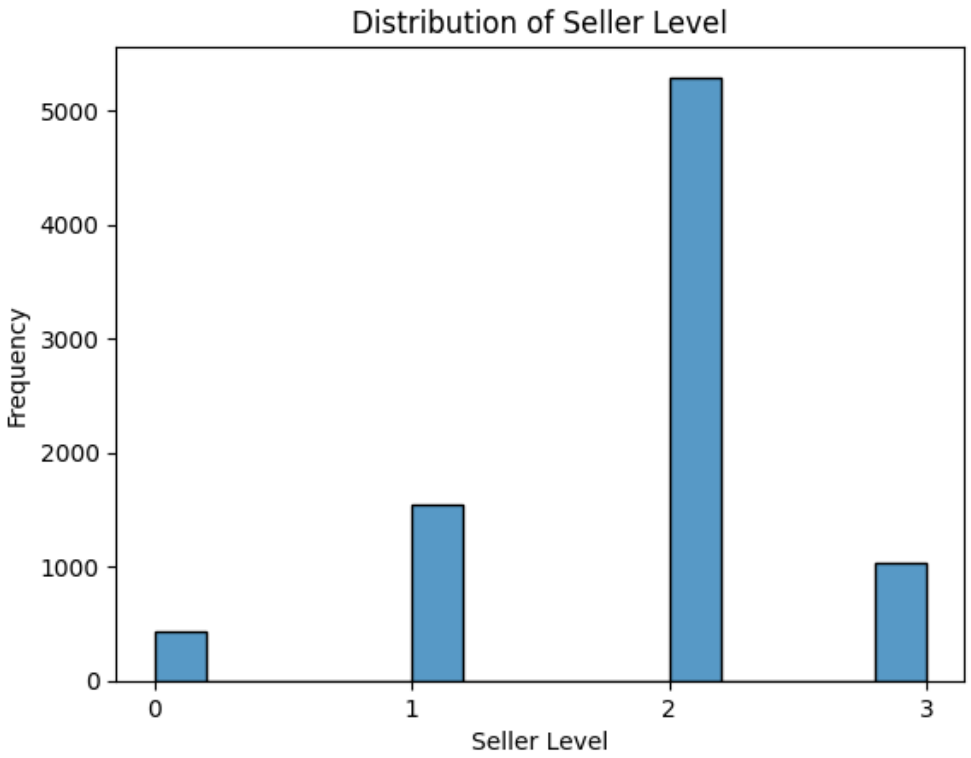
20 Most Common Words in Tags

השפעות על הציון בצורה ויזואלית

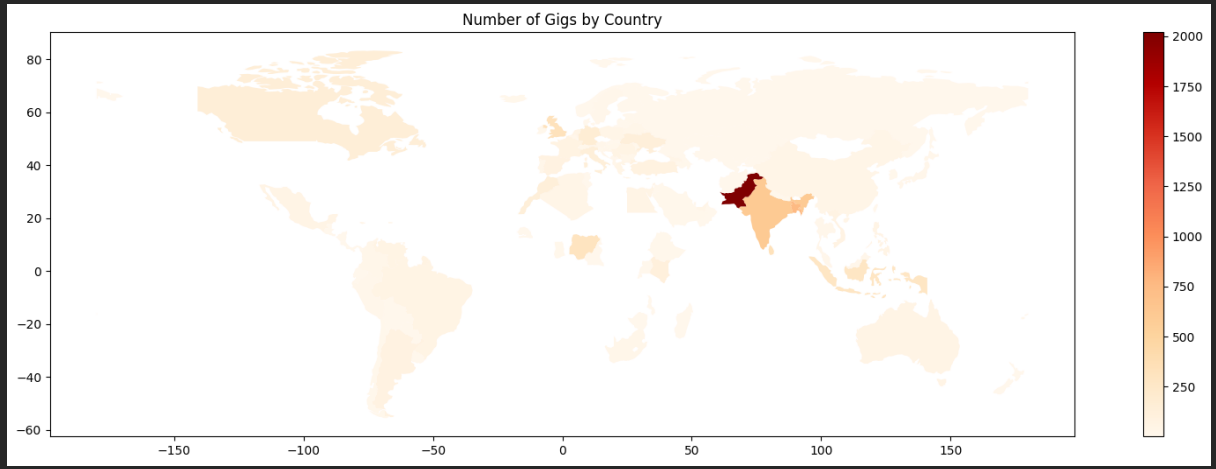


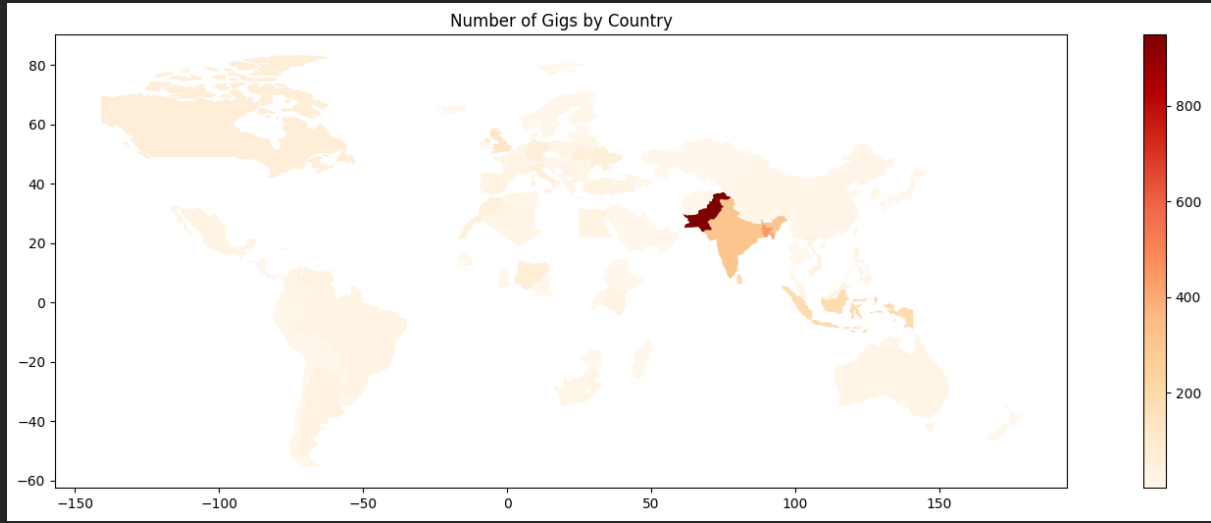
החלוקה של משתנים





הגיגים שנוצרים בעולם



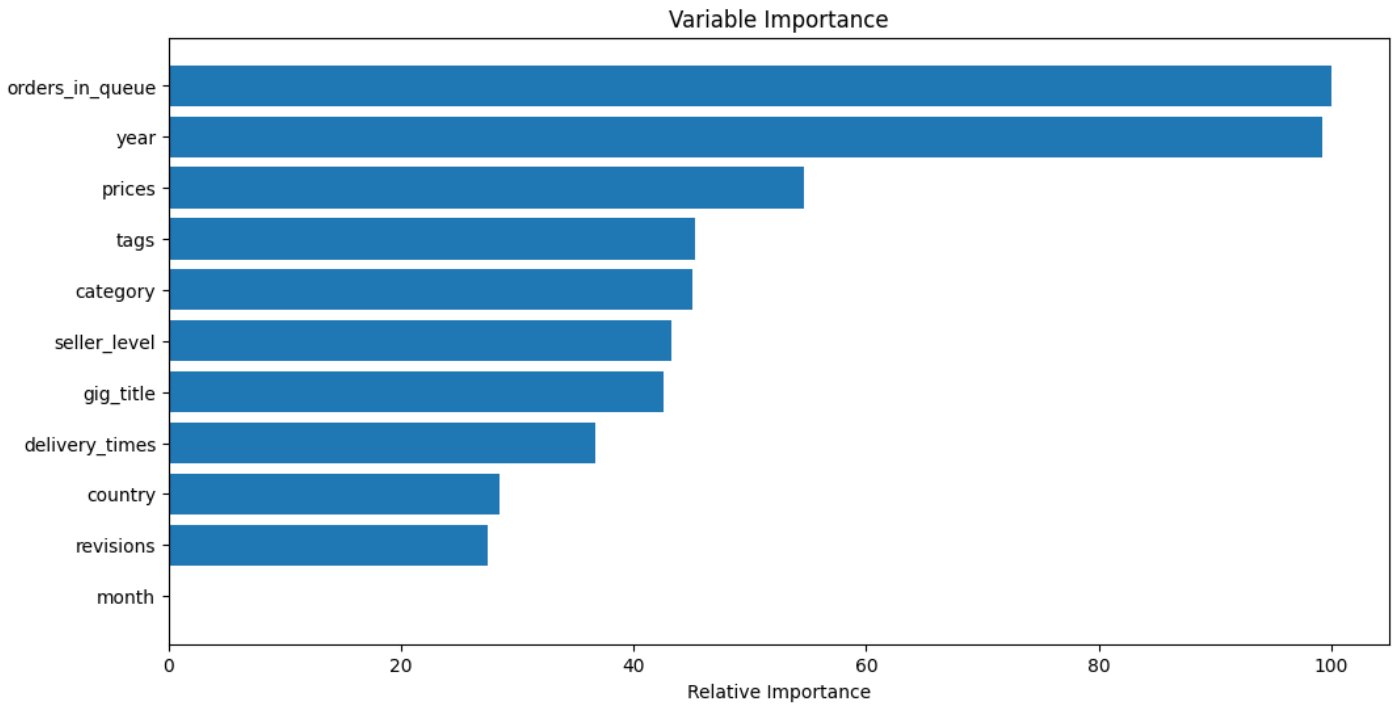


בתמונה למעלה: כל הגיגים

בתמונה למטה: רק גיגים מצליחים - מעל החציון

בתכלס - אותו דבר. אסיה בקלות ובעיקר - פקיסטן שולטת בכמעט כל פייבר.

יישום הפתרון ובדיקה של השיטות שיושמו (הערכת ביצועים)



ניסינו כמה שיטות כדי לפתור את הבעיה. בגלל כל המשתנים הקטגורליים שלנו והרבה מחרוזות, היה לנו קשה להשיג מידע מספרי לבצע חישובים.

ניסיון של עץ החלטה הביא אותנו לדיוק של 30% (בבדיקה על הtest data) בחיזוי הscore המדוייק. (לא רע בהתחשב בscore).

ולגבי שאלת המחקר ה"אמיתית" שלנו - > האם גיג מצליח? השתמשנו ביער אקראי, ואחרי אופטימזציה, הצלחנו להגיע לדיוק של 85.7% שיגיד לנו האם gig מצליח (כאשר אנו מגדירים מצליח כמעל הממוצע). בצד ניתן לראות "סיכום" של השפעה של כל משתנה.

הסקת מסקנות ודיווח מסכם

- לכותרת יש השפעה על הצלחת הgig והמילים שתשתמשו בו יכולים להשפיע על הצלחתם.
- אנשים אוהבים מחירים זולים, וזמן העברה קצר, ככל שתהיו יותר זולים ותכינו את העבודה מהר, באופן צפוי, תצליחו יותר.
- לשנה שאתם מצטרפים יש משמעות "מינימלית" להצלחה -> אבל להיות משתמש חדש ייפגע מאוד בסיכויים שלכם להצליח
- רמה גבוהה בפייבר תשפיע משמועתית על הסיכויים שלכם להצליח (מה שמוכיח שההזמנות הראשונות הן הכי קשות - ואז זה יכול להיות הרבה יותר קל).

בגדול - הצלחנו לפתור את הבעיה שלנו. אנחנו יכולים לחזות האם גיג יהיה מצליח, ואפילו בקירוב גם כמה הוא יצליח, עוד מלפני שהוא הועלה לאינטרנט. רק על פי נתוני המשתמש והגיג עצמו.

שימוש במודל שלנו יאפשר להגדיל את הסיכויים שלכם להצליח, והתצוגות והויזואלזציות שהכנו מאפשרות לדעת בדיוק מה אנחנו צריכים לעשות ולשנות כדי להצליח באתר התחרותי הזה.