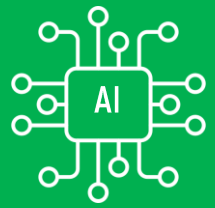
Advertisement Campaign Analysis

Anna Yordanova, Data Scientist

27.08.2025

Analyze



Empower

Implement

Grow

Data Overview

Socio-Demographics

15854 rows

- weight
- household
- birthday
- children
- gender
- person_id

Purchase - our target feature

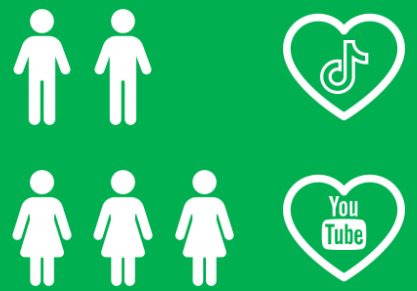
- purchase
- channels **x13**

TV
Flyers
Print
Online x2 + total
TikTok
YouTube x4 + total
Pinterest

Media Contacts

15854 rows

Weight – number of people with identical features



Each row represents a large group of people

We should try to preserve as many rows as possible

Data Cleaning

1. Extract features:
   1) '**birthday**' – '**age**'
   2) '**household**' – '**gender_temp**'
   Delete '**birthday**'
2. Extract numeric data from '**children**' and '**household**'; transform '**gender**' to integer
3. Find bad values in '**weight**' (17%); try regression model; fill with 1
4. Fill missing values in '**gender**' using info from '**household**' + weighted mode
5. Fill missing values in '**children**' and '**channels**' with 0
6. Fill missing values in '**age**' and '**household**' using weighted mean
7. Check sums in the columns '**x_total**'

Merge Datasets on index & person_id

8. Check duplicates without '**weight**'
9. Aggregate rows with identical features and sum up their '**weight**'
10. Check the dataset for features with constant values
11. Make sure total '**weight**' before = total '**weight**' after

The data is clean and ready!

ML Model Development

Preprocessing

- Train/Test Split (70/30, stratified)
- StandardScaler for numeric non-binary features
- No categorical features to encode
- Two sets for multicollinear features

Cross-validation

- StratifiedKFold (5 splits)
- Models:
  - LogisticRegression x2
  - kNN
  - SMV
  - XGBoost
  - LightGBM
  - CatBoost

Tuning

- Optuna
- CatBoost:
  - learning_rate
  - l2_leaf_reg
  - random_strength
  - max_depth
- LogisticRegression:
  - C
  - penalty
  - solver

  Set: no total features

Logistic Regression

C = 0.00047694
penalty = "l2"
solver = "saga"
random state = 42

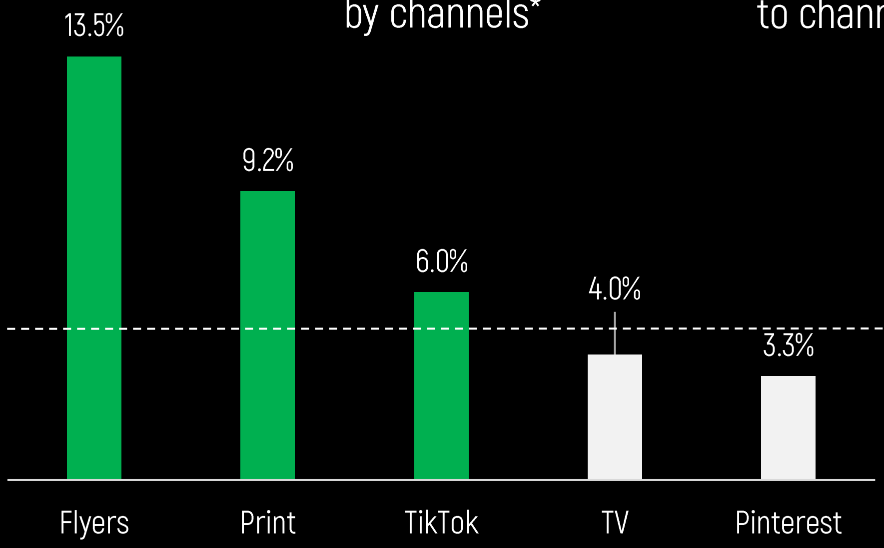
	F1	Precision	Recall
CV	0.734	0.614	0.913
Test	0.729	0.634	0.857

No overfitting
Good result

Conversion Rate

= purchases / population exposed to channel

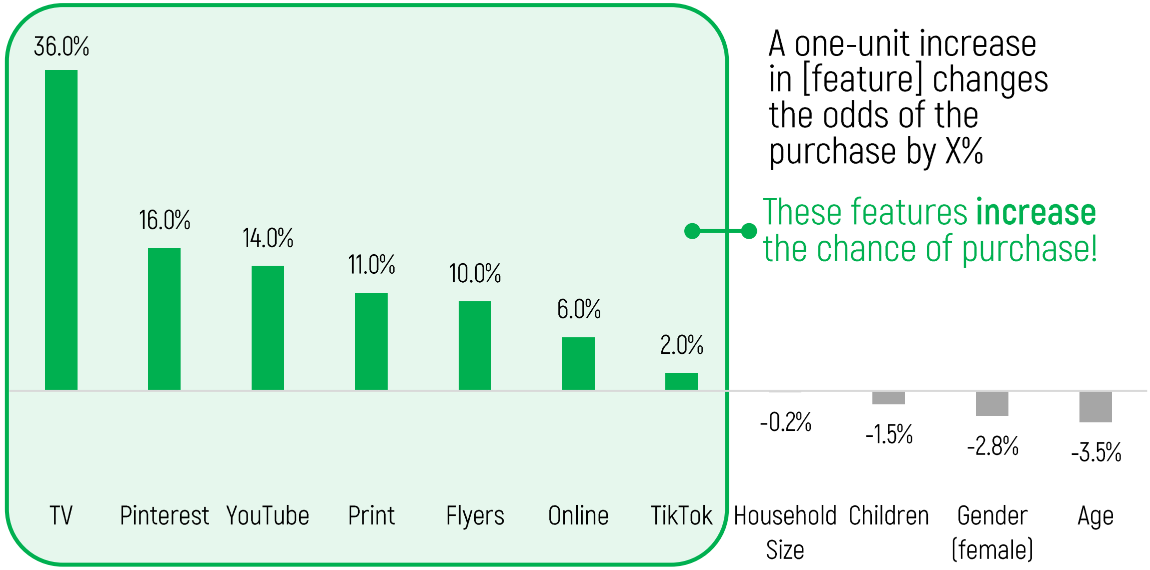
by channels*



5.11% Overall Conversion Rate

* only standalone channels

ML Model Odds Ratios



A one-unit increase in [feature] changes the odds of the purchase by X%

These features **increase** the chance of purchase!

Customers

9% Organic

<40 years old
no children
lives with partner

91% Acquired

41+ years old
no children
lives with partner

Key Points

Leverage Flyers

High CR (13.5%), especially effective for the 71+ age group

Utilize TikTok

High CR (6%) but underused (0.47% of all interactions)

Prioritize TV

Dominates interactions (63%) and boosts purchase odds by 36%

Focus on Male Customers

Data and model indicate higher conversion likelihood for men

Would you like to see Jupyter Notebook with this research?

Click here: https://link.example.com

Have a nice day!