Predict a Click!

trivago Case Study

Carlo Contaldi

Data Science Workflow

- Frame the problem
- Exploratory Data Analysis
  - Missing data
  - ID characterization and uniqueness
  - In-depth feature analysis
  - Regression Analysis
  - Data representativeness
- Production Pipeline
  - Data preparation & feature engineering
  - Preprocessing
  - Training, Validation & Test

Frame the Problem

trivago – tech company providing lodging meta search services

Regression task – predict `n_clicks` of given hotel entry

- Evaluation metric – WMSE

$$wmse := \frac{1}{n}\frac{\sum_{i=0}^{n} w_i \cdot (predictedClicks_i - observedClicks_i)^2}{\sum_{i=0}^{n} w_i}$$

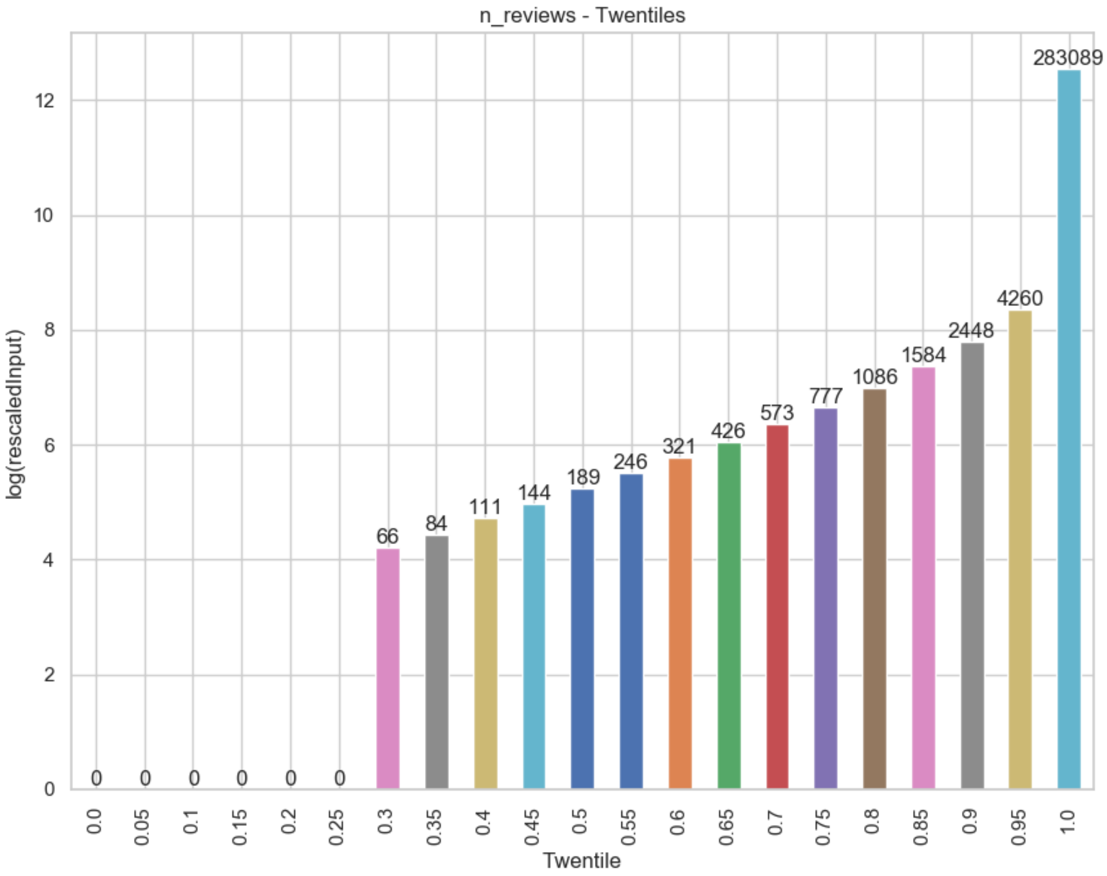
$$w_i := log(observedClicks_i + 1) + 1$$

- ~400K entries
  - Features – `hotel_id`, `city_id`, `content_score`, `n_images`, `stars`, `distance_to_center`, `avg_rating`, `n_reviews`, `avg_rank`, `avg_price`, `avg_saving_percent`, `n_clicks`
- Conda 4.5.11 with Python 3.6.5 on Win10 x64
  - NumPy, Pandas, Matplotlib, Seaborn, Scikit-Learn, XGBoost

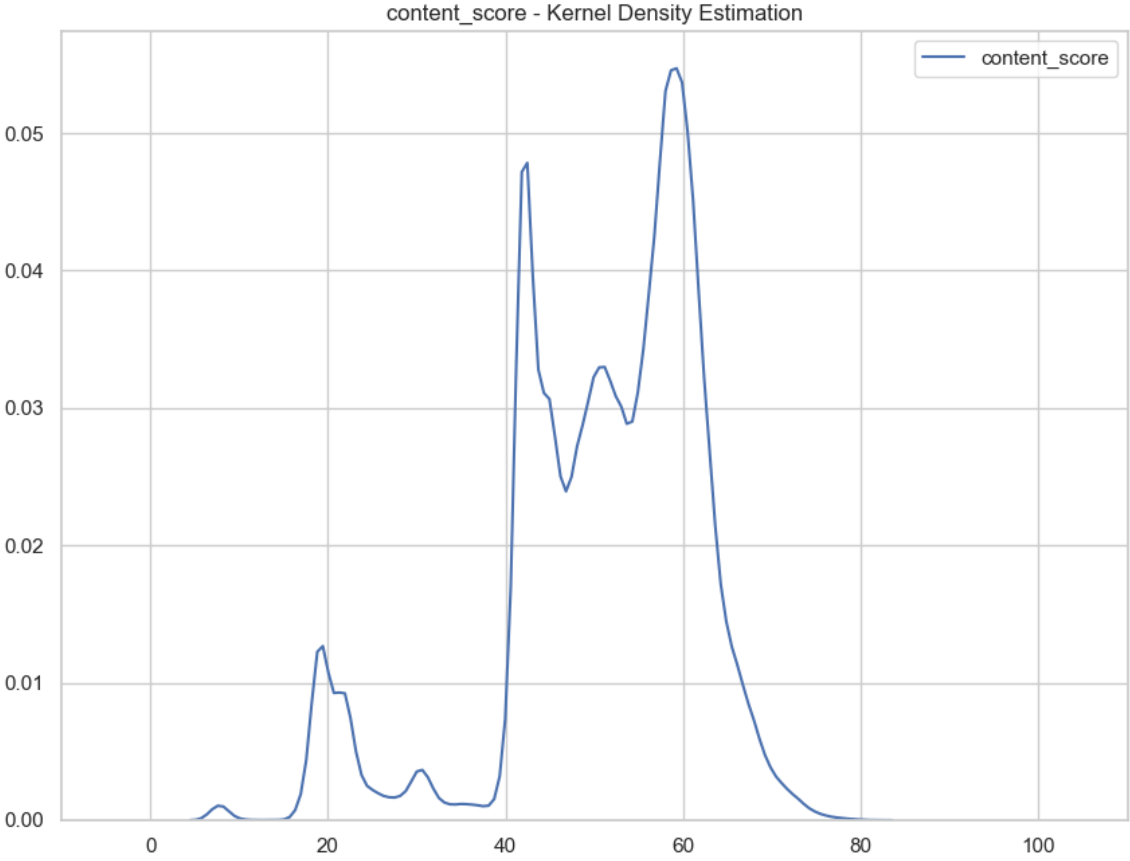
Exploratory Data Analysis (I)

- All entries have distinct `hotel_id`
- # missing values negligible excepting for `avg_rating`
  - `n_reviews=0` ➔ `avg_rating=NaN` (Cold Start problem)
    - Naïve mean imputation / Regression imputation
- `city_id` – ~30k categories
  - One-hot encoding + shrinkage / clusterization
- ~1% of data has `n_images=-1`
- Insights from quantile/KDE/box plots
  - Extremely skewed distributions ➔ `feature = logp1(feature)`
    - `n_clicks`, `n_images`, `distance_to_center`, `n_reviews`
  - All features have reasonable distributions

Exploratory Data Analysis (II)

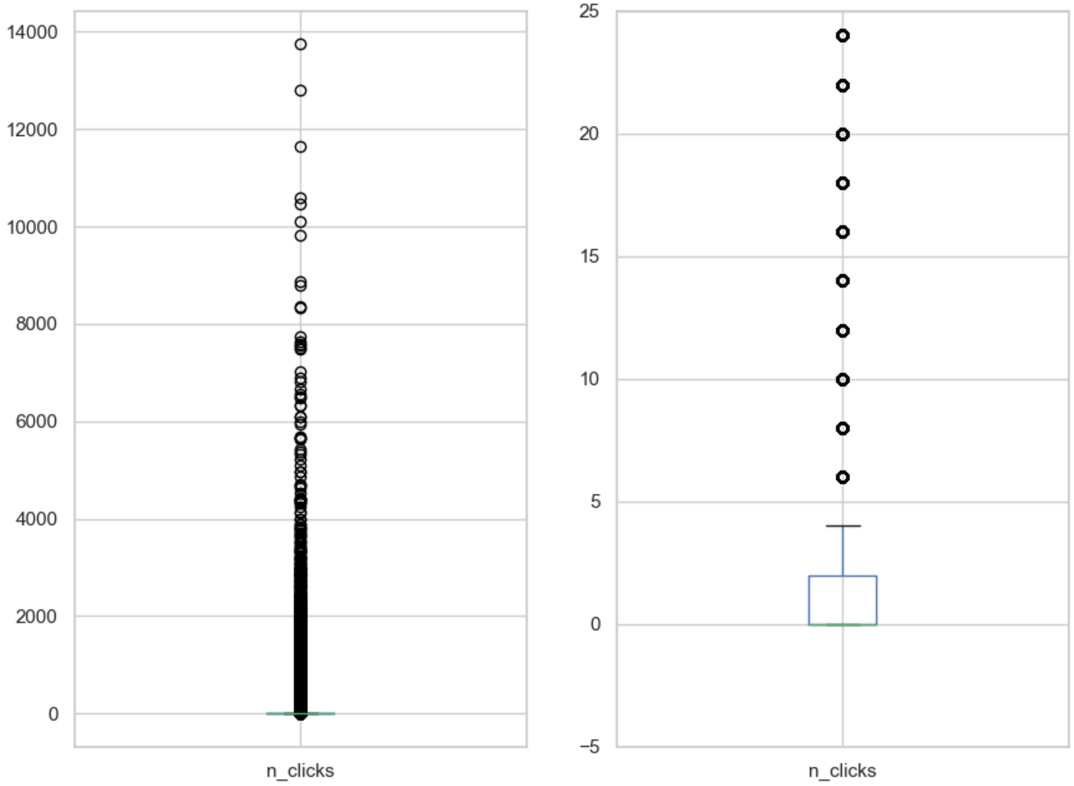


Exploratory Data Analysis (III)



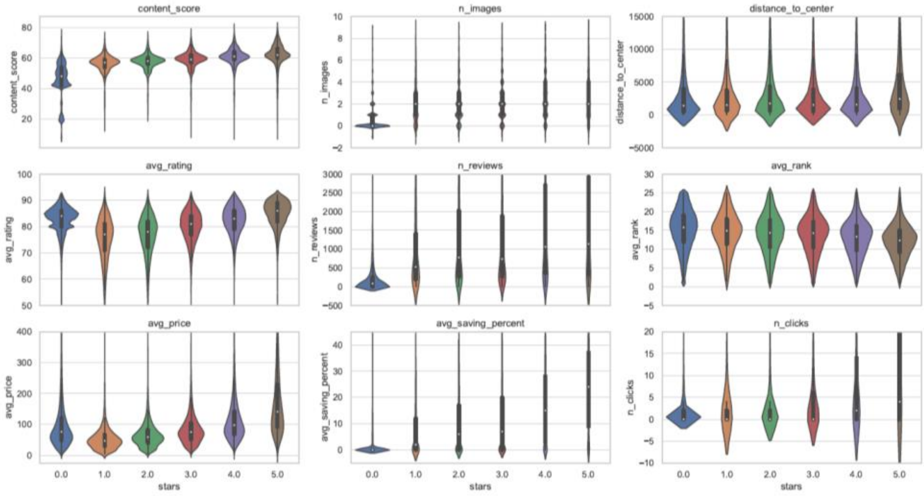
Exploratory Data Analysis (IV)

n_clicks - Box Plots



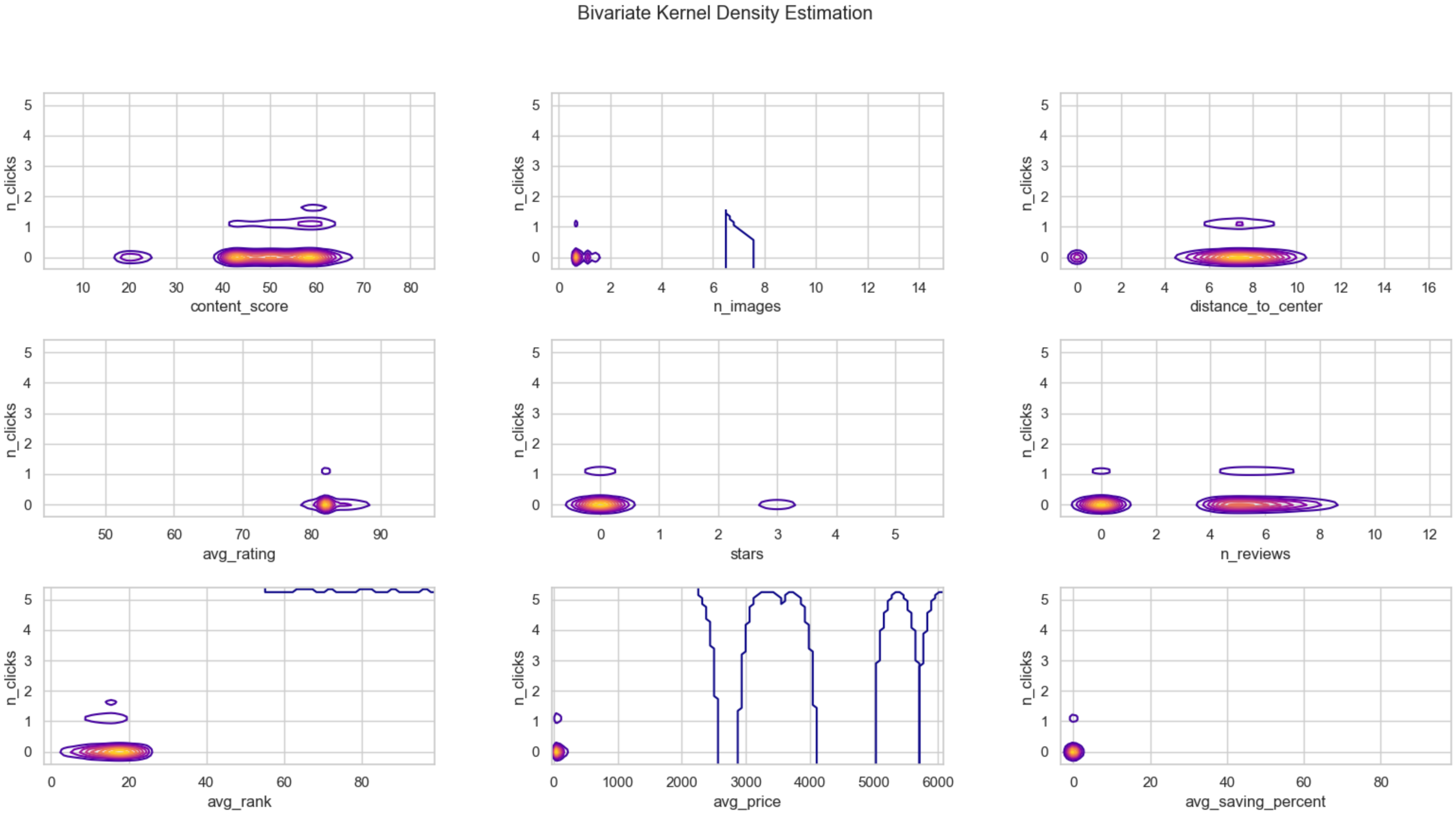
EDA – `stars`: the unknown feature

- Domain: integer in [0, 5]
- What does it describe?
  - Violin plots ➔ hotel stars rating
    - `content_score`, `avg_price`, `avg_rating` grow with `stars`
    - `avg_rank` decreases with `stars`

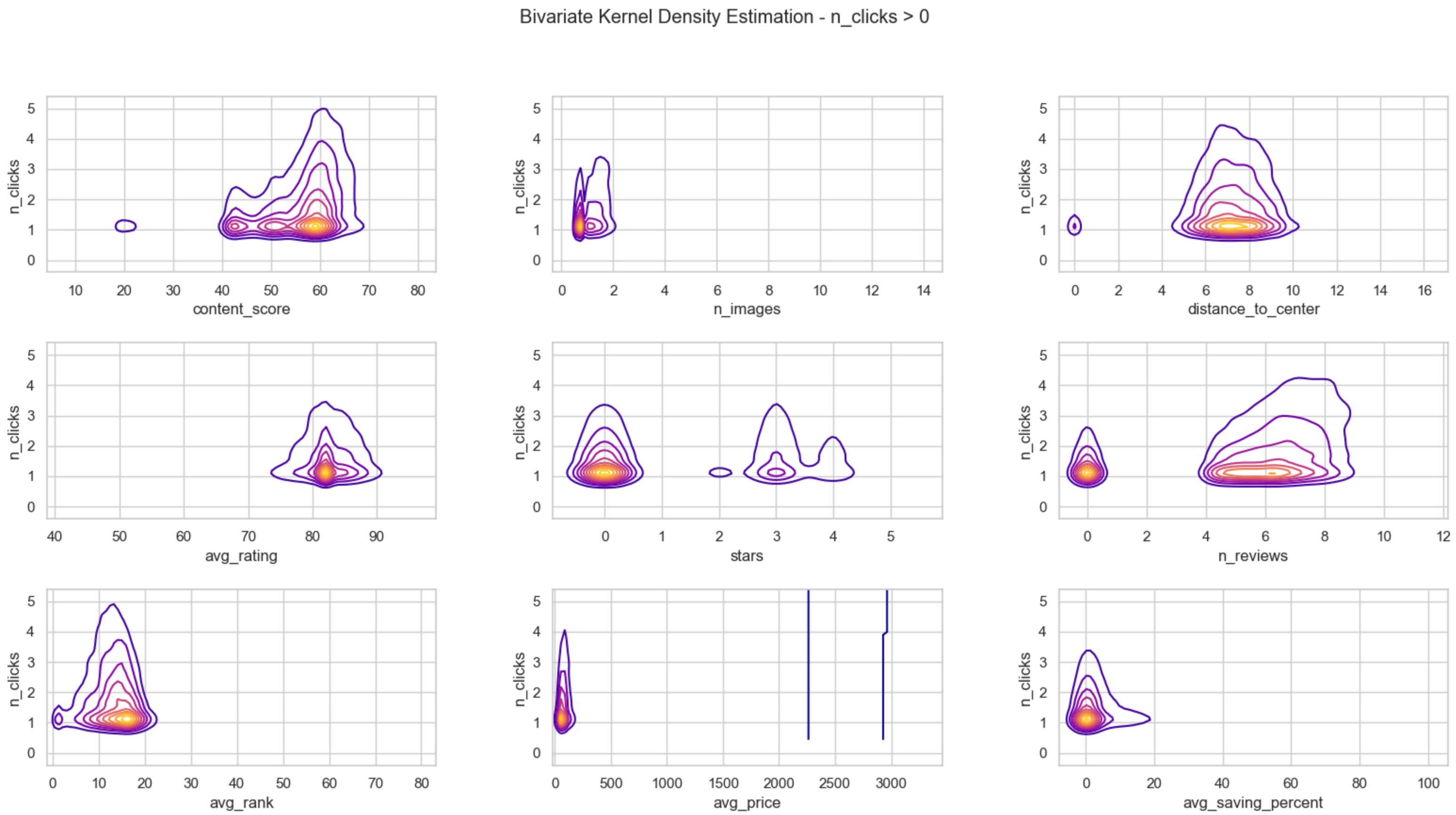


- What is a 0-star hotel?
  - Violin plots & trivago website ➔ hostels & aparthotels
    - Consistently lower `content_score`, `n_reviews`
    - Most of 0-star entries have 0 `n_images` and `avg_saving_percent`
- Solutions: categorical / numerical+regression imputation

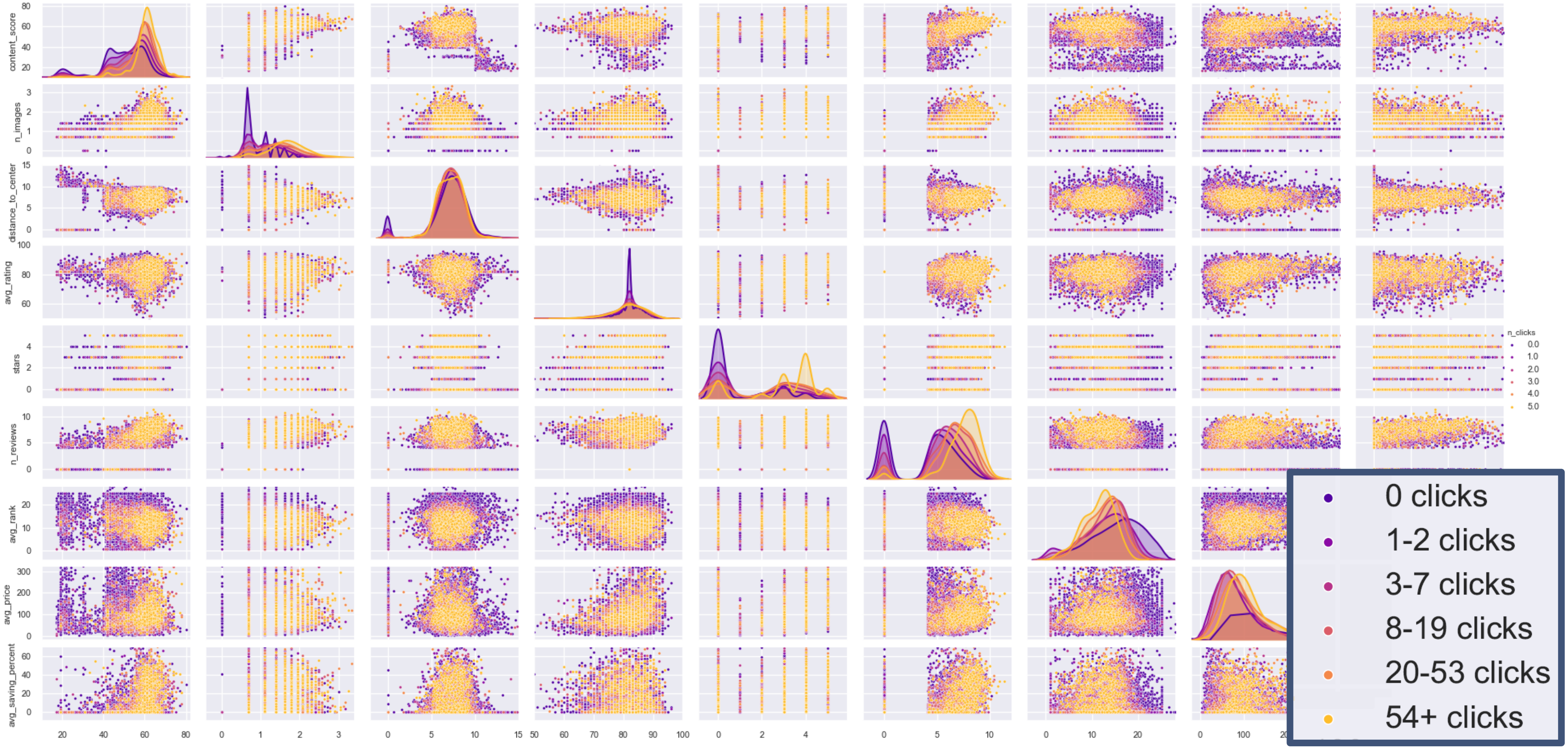
EDA – Bivariate KDE



EDA – Bivariate KDE (n_clicks>0)



EDA – log-Discretized Pairplots



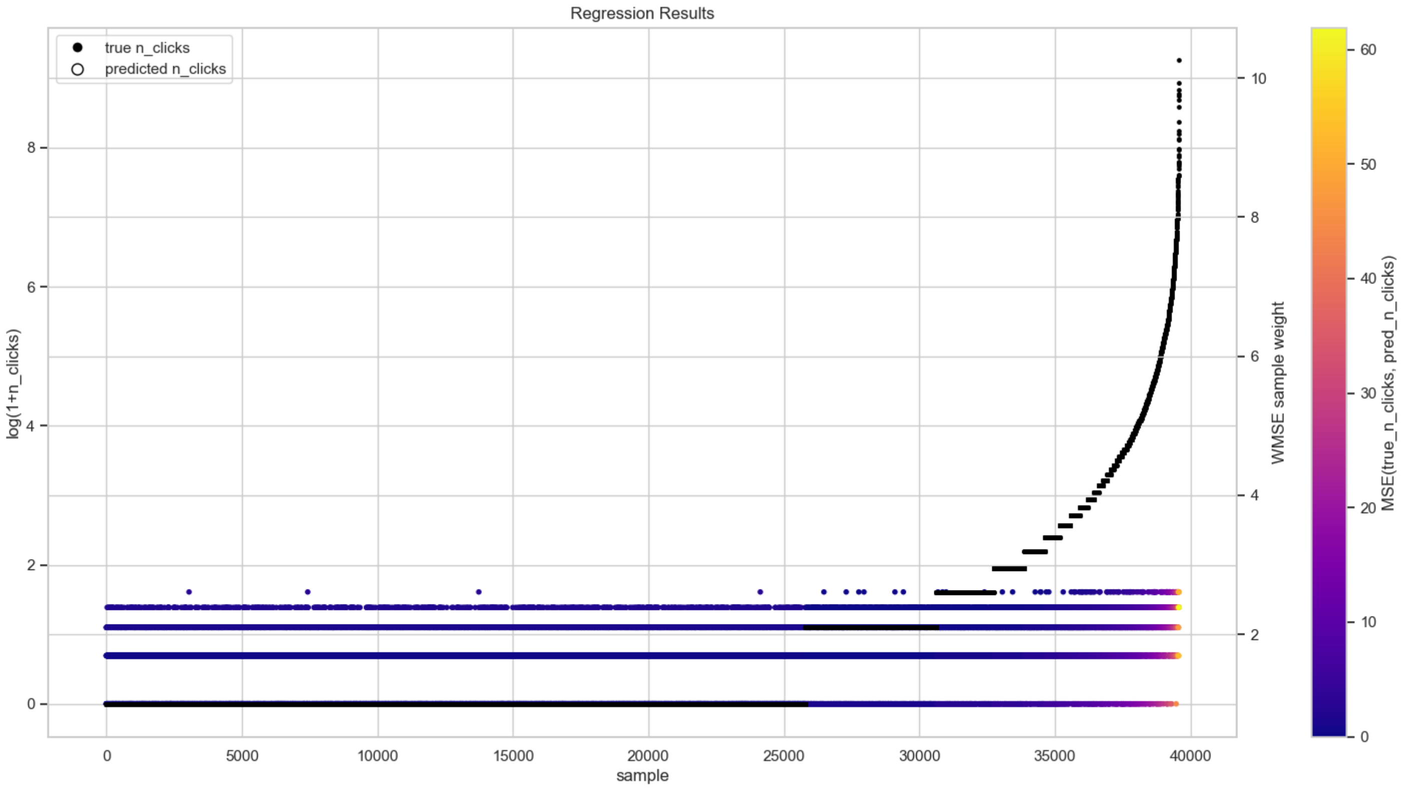
Experimental Framework

- Training & Validation / Test – 90/10 randomized split
- Experimental reproducibility (`random_state=0`)
- Preprocessing
  - Production-ready: type cast, NaN dropping/handling, domain checks
  - Min Max Scaler / Max Abs Scaler
  - One-Hot Encoding / Truncated SVD
- Training
  - 5-fold randomized Cross-Validation
  - Grid search / Randomized search
  - XGBoost with WMSE-based Early Stopping
  - Statistical significance assessment: Wilcoxon Test

ElasticNet, PolyRegression, SVR

WMSE ~= 2.24

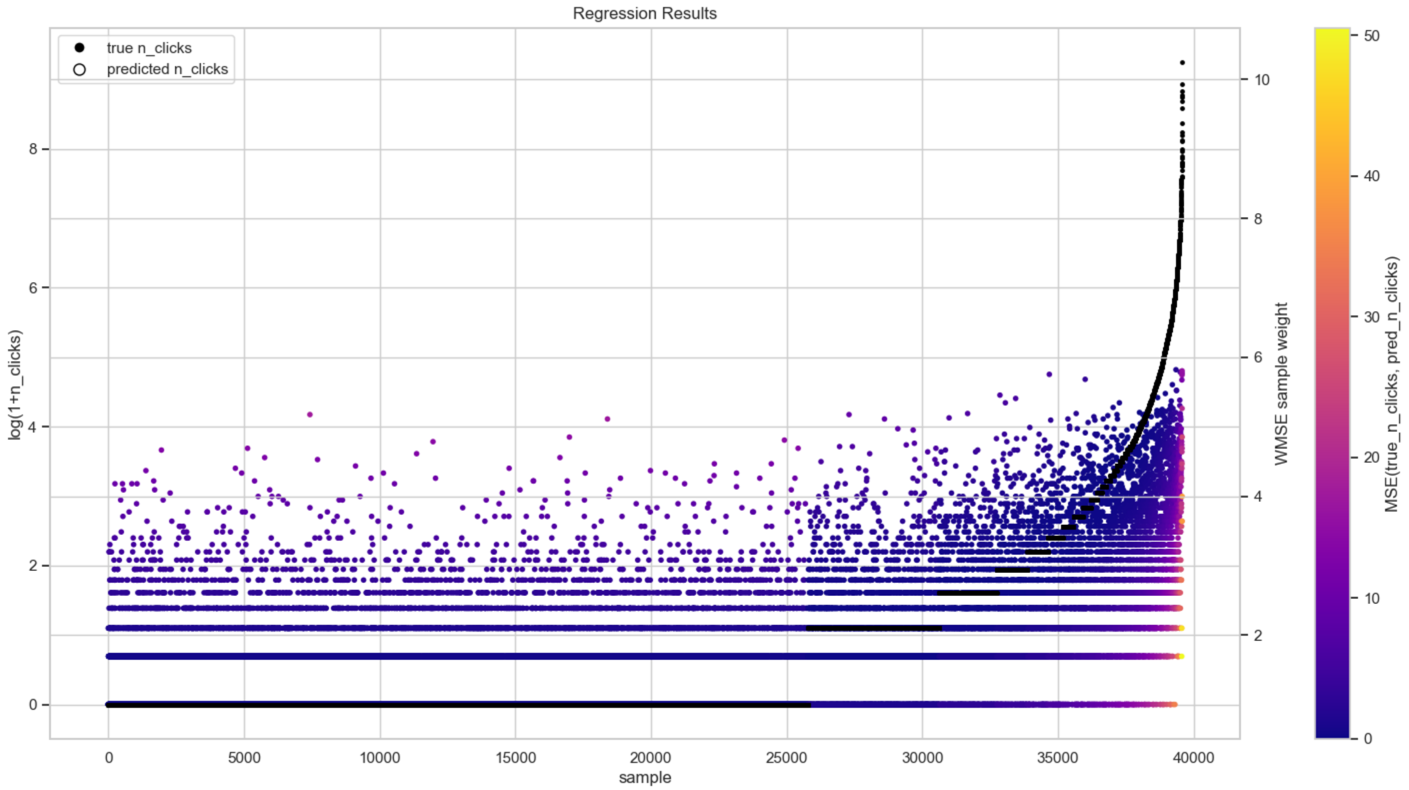
Random Baseline WMSE = 2.24



Random Forest

WMSE = 2.20

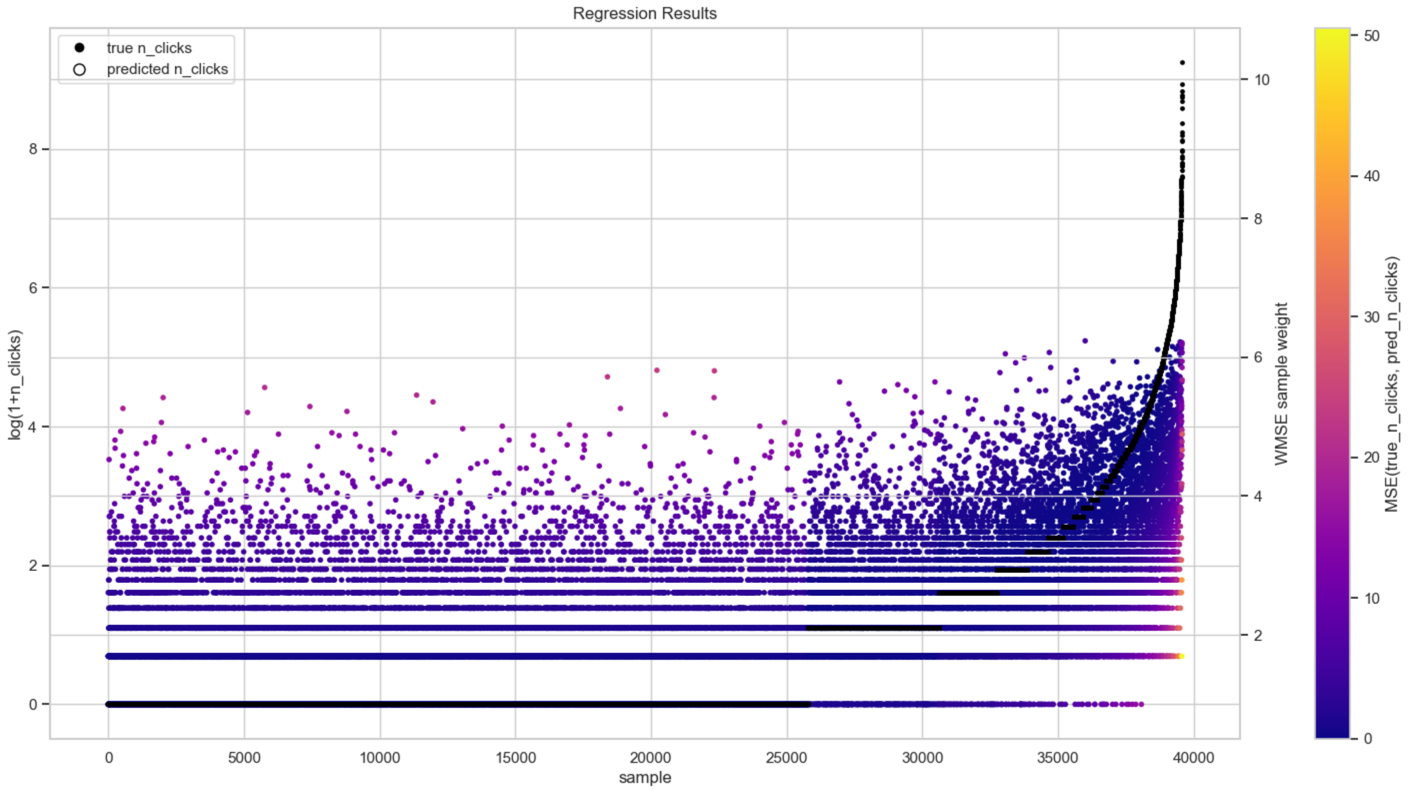
Random Baseline WMSE = 2.24



Random Forest + Weighted Oversampling

WMSE = 2.12

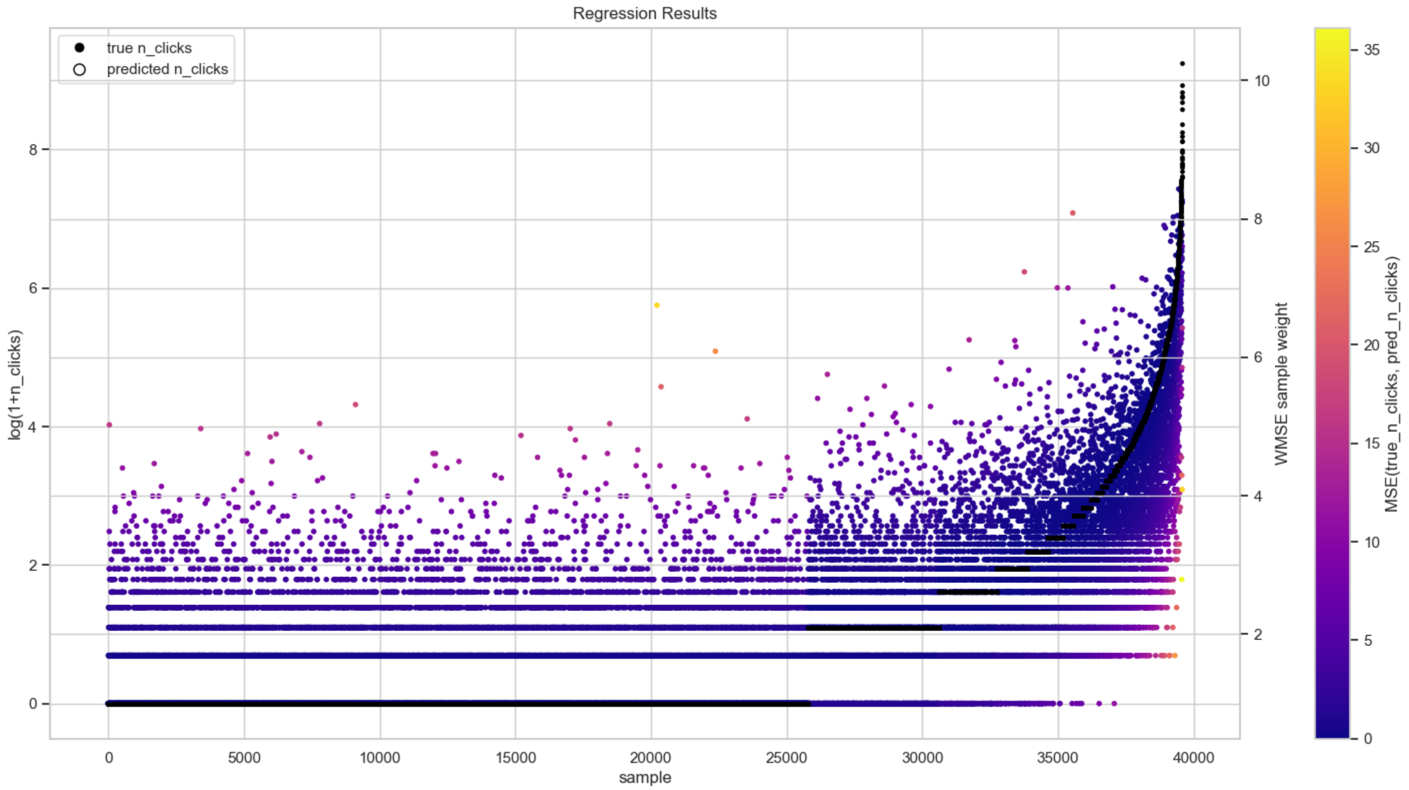
Random Baseline WMSE = 2.24



XGBoost + Weighted Oversampling

WMSE = 1.66

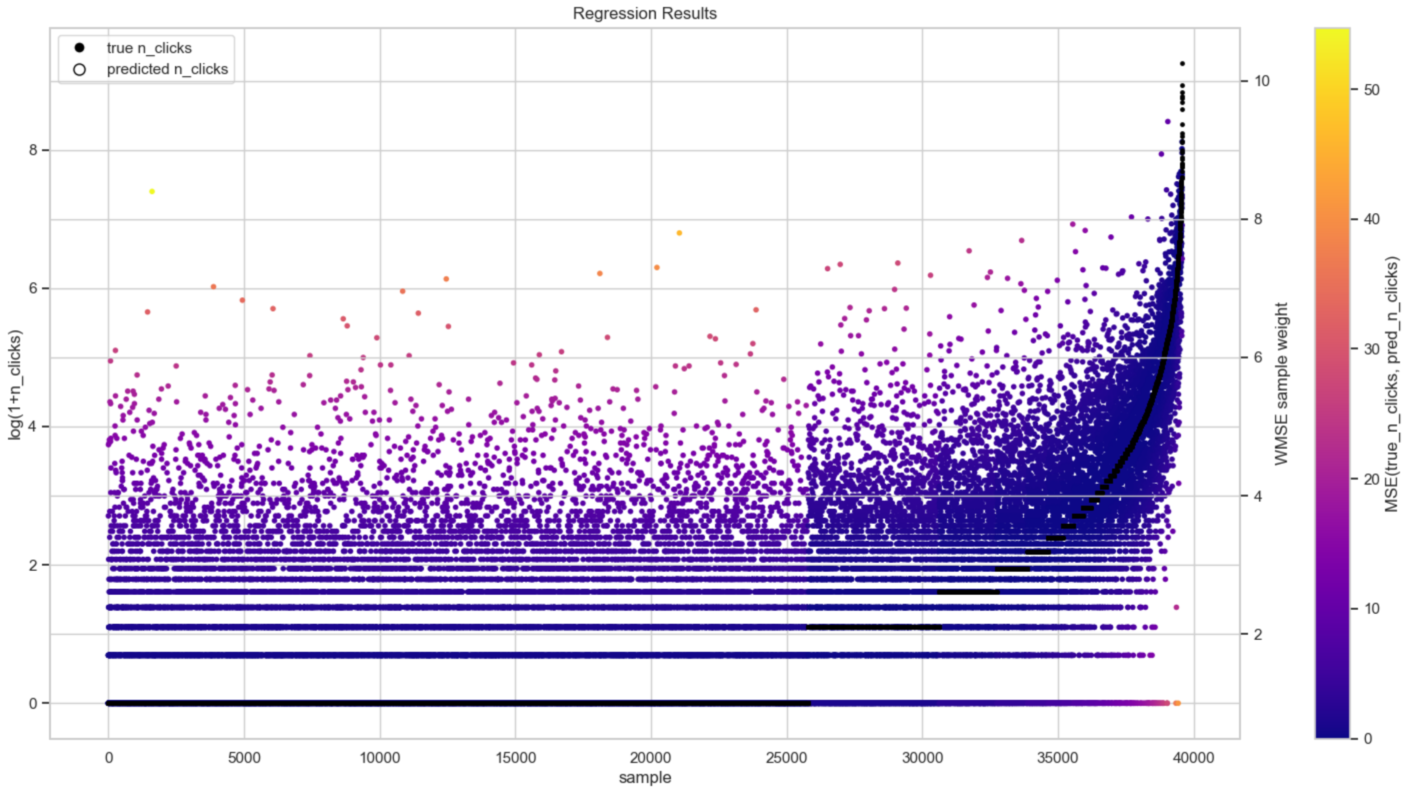
Without WO, WMSE = 1.74



XGBoost + Weighted Oversampling, no log

WMSE = 1.22

Without WO, WMSE = 1.26



XGBoost + WO, no log, stratified splitting

WMSE = 0.91