Data Science - Growth team

Rappi

October 10, 2019

María Fernanda Cortés Soler

Table of contents

1. Overview
2. Conventional response models
3. Uplift modeling
4. Methodology
5. Conclusions

Our Growth Team

[Our goal]

Lead users to make an extra purchase within a determined period of time

[Strategy]

Customer-facing actions such as:

- Monetary incentives
- Offers
- Advertisement

[Business decision]

Which users should we contact with these strategies?

[Key points]

- Generate incremental impact (additional purchases)
- Reducing costs
- Do not unnecessarily annoy customers

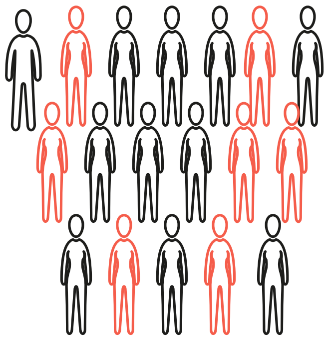
Conventional response models

Predicting the ‘good’ and the ‘bad’ customers

- Build a classifier to figure out which customers are most likely to buy after receiving a treatment
- These customers are selected as target

Weaknesses:

- Unnecessary costs
- There are users so prone to purchasing, that will do so even if not contacted
- Model designed to maximize the response rate, instead of the incremental impact



Which red respondents would have purchased anyway?

Uplift modeling

We should only target users that are likely to buy if contacted by the campaign, but unlikely to buy otherwise

Buy if do receive treatment			
	No	Do not disturb	Lost causes
	Yes	Sure users	Persuadables
		Yes	No

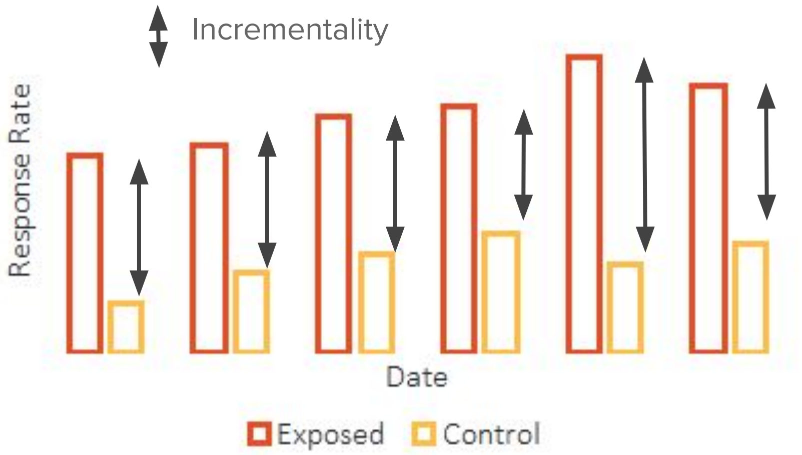
Buy if don't receive treatment

It's only concerned with incremental effect of a treatment in the population.

Population

Control

Exposed



Methodology - Uplift Modeling

User profile x

Two Machine Learning models:

Incentive model
P(response | offer, x)
P_1

No incentive model
P(responsel no offer, x)
P_2

Response variable:

Conversion
Customer purchases in the current date.

Uplift =
P(response | offer, x) -
P(responsel no offer, x)

Methodology - Uplift Modeling

Dataset variables:

User's history

- User discounted orders
- Average number of products
- User finished orders
- Time from registration to 1st order
- User average total value
- User average tip
- Acquisition channel
- User number of CC usages
- User cancelled orders
- User number of verticals
- Number of orders during week
- Number of orders in weekend
- Average discount proportion
- Last order hour

Segmentation

- User latitude-longitude score
- Device score (price-year)

Recency

- Days since last order

App events

- App launches of last 7 days

Methodology - Uplift Modeling

Dataset records:

- 653,244 customers
- Purchase history in past six months
- Monetary incentives related data
- Resampling technique: SMOTE *(imblearn)*

We experimented with two different classifiers:

- SVM
- Random Forest (*sklearn*)
- **XGBoost (*XGBoost*)**

User id	Days since last order	...	User variables	...	Granted incentive	Purchase
12345	1		...		1	0
12345	2		...		1	0
12345	3		...		0	0
12345	4		...		1	0
67891	5		...		0	1
34567	6		...		1	0
34567	7		...		0	1

- Incentive model
- No incentive model

Uplift Modeling - Results

Hyperparameter tunning:

- Randomized Search (sklearn)
- 3-fold cross validation

Uplift =
P(response | offer, x) -
P(response| no offer, x)

⚠ **We cannot directly substract the scores of both models.**

Brasil

ROC Area under the curve

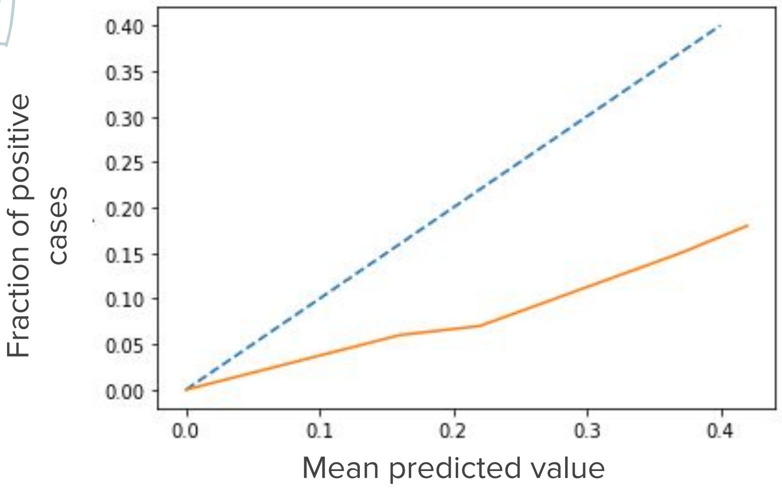
Incentive model	No incentive model
0.816	**0.741**

- Not calibrated probabilities.
- We need predicted probabilities that match the expected distribution of probabilities for each class

Uplift Modeling - Results

No incentive model: P(response | no offer, x)

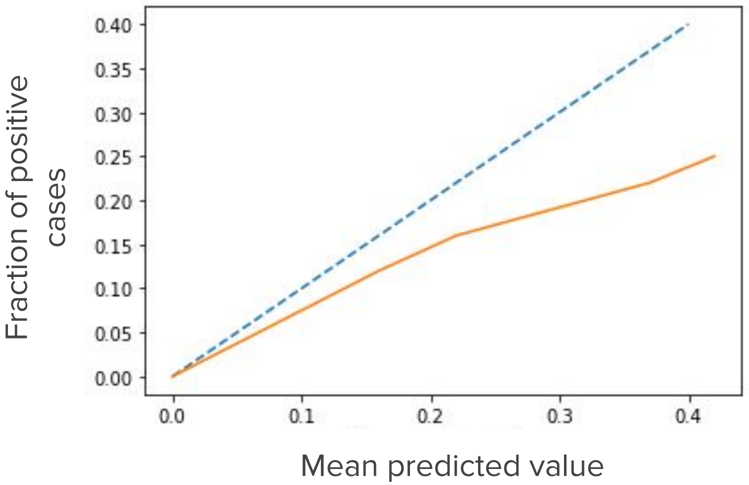
Reliability diagram:



The observed fraction of positive cases does not match distribution of predicted probabilities (BR).

Calibrated model: Isotonic regression

Reliability diagram:



The better calibrated, the closer the points will appear along the main diagonal (BR).

Uplift Modeling - Results

Calibration methods:

Isotonic Regression:

- Fits a non-decreasing function to a vector of elements
- It forms a function that is piecewise linear.

Computation time

Fits the model using k-fold cross-validation
Calibrate the probabilities predicted by these models

Libraries: *sklearn.calibration*

```
from sklearn.calibration import calibration_curve
from sklearn.calibration import CalibratedClassifierCV
```

Prepare Data

```
train, test = train_test_split(df, test_size=0.5,
                               stratify=df[label],random_state=23)
```

Define base model

```
model=XGBClassifier(booster='gbtree', colsample_bylevel=1,
              colsample_bynode=1, colsample_bytree=0.6, gamma=0,
              learning_rate=0.01, max_delta_step=0, max_depth=5,
              min_child_weight=2, missing=nan, n_estimators=300,
              objective='binary:logistic', subsample=0.8, verbosity=1)
```

Fit and calibrate model on training data

```
calibrated = CalibratedClassifierCV(model, method='isotonic', cv=5)
calibrated.fit(train[features], train[label])
```

Methodology - Uplift Modeling

User profile x

Two Machine Learning models:

Incentive model

P(response | offer, x)

P_1

No incentive model

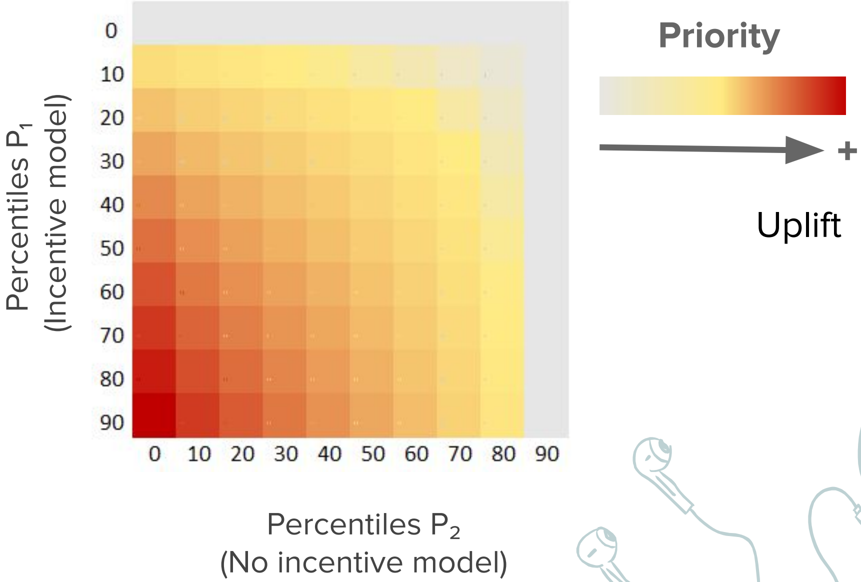
P(responsel no offer, x)

P_2

Response variable:

Conversion

Customer purchases in the current date.



Methodology - users to select

[Our aim]

Estimate the inorganic GMV (paying for ads, etc) needed to achieve the monthly goal for momentum users

[In-month inorganic income needed]

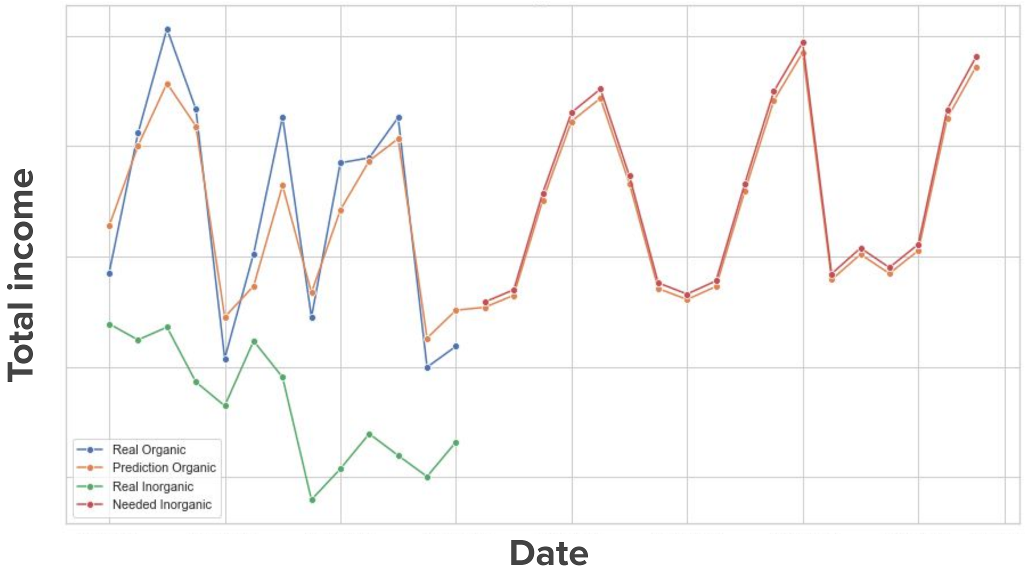
=

Monthly income goal

-

predicted in-month organic income

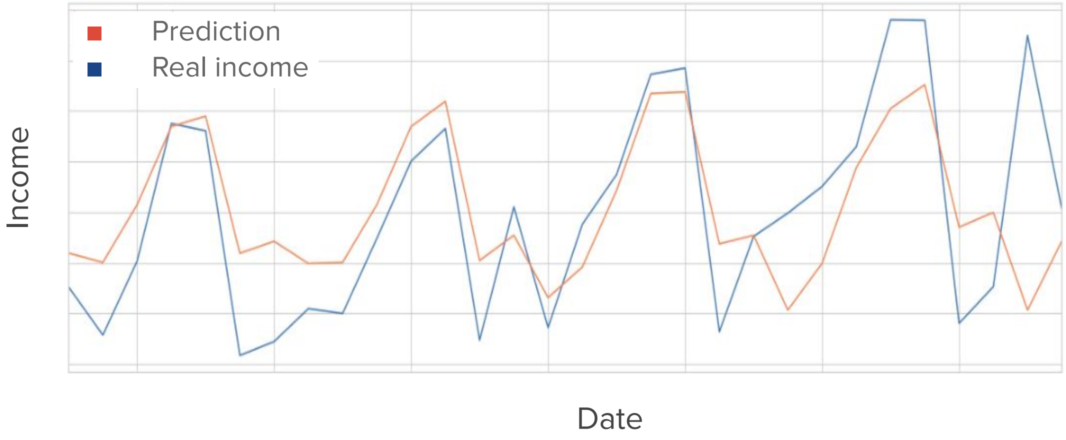
Total monthly income by day



- Future in-month inorganic income needed
- Previous in-month organic income
- Predicted future in-month organic income
- Previous in-month inorganic income

Methodology - users to select

Predicting future organic income



Key observations:

- Weekly and monthly seasonality
- Special events (Holidays)
- Positive trend

[Methodology]

XGBoost regression model:

- Allows needed features (holidays)
- Daily data since Jan 2019
- Hyperparameter tunning: Grid Search

Monthly relative error

4.15%

Methodology - overall process

Rappi

first

Prediction of organic in-month income

second

Estimate in-month required inorganic income

third

Estimate needed number of users to contact

last

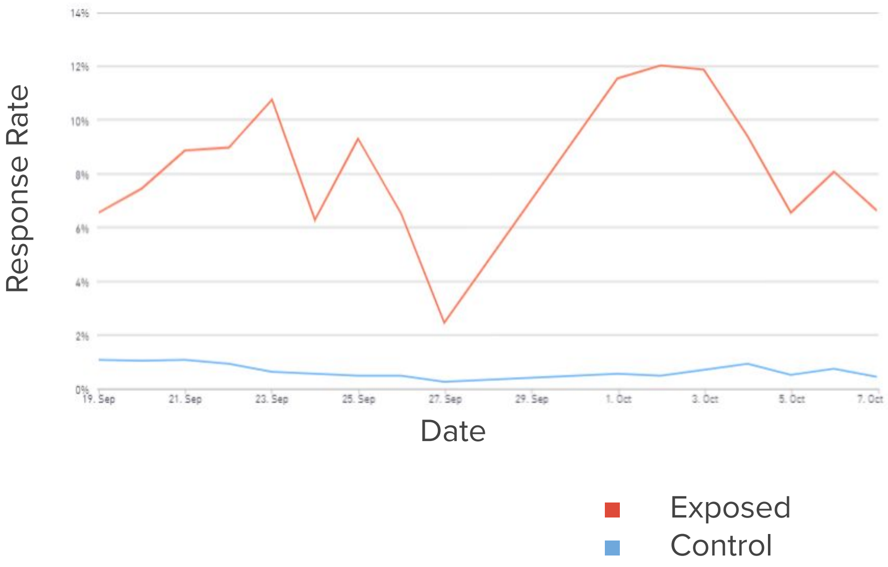
Select users with uplift model

Conclusions

- Our experiment confirms the usefulness of uplift modeling in marketing campaigns
  - Positive overall uplift rate: Differences in conversion for users contacted by the campaign and users who didn't receive it.

- Uplift modelling allows us to better target the intended set of users.

Brasil results



References

- Siegel, E. (2011). Uplift modeling: Predictive analytics can't optimize marketing decisions without it. Prediction Impact white paper sponsored by Pitney Bowes Business Insight.
- Kumar, A., & Kumar, R. (2018). Uplift Modeling: Predicting incremental gains.

Rappi

Thank you!

Any questions?

You can find me at:
maria.cortes@rappi.com