Data Scientist Interview- MBM

XIAOCUI(JELLY) ZHANG
2020-08-03

Content Layout

- Tasks Introduction
- Solution and results of Task1
- Solution and findings of Task2
- Summary, questions and extensions

Tasks Introduction

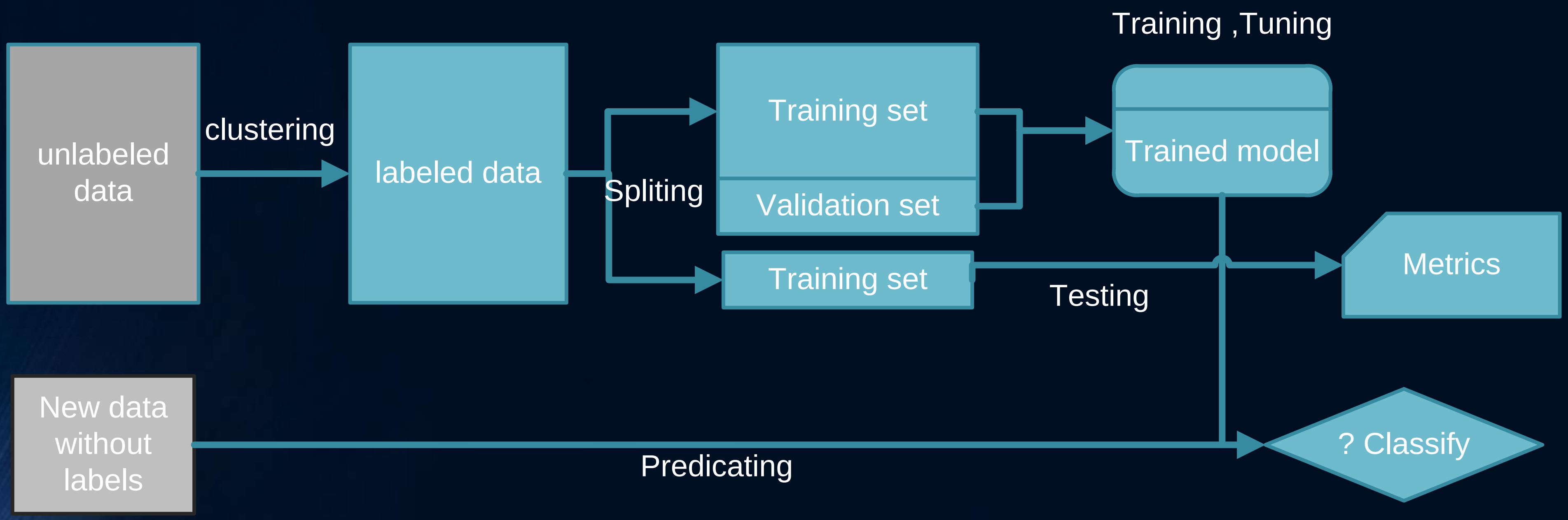
ONLINE POTENTIALLY BAD CONSUMERS IDENTIFICATION

1. Data
   - Account Features
   - Personal ID Features
   - Redemption Activity
2. Key technical points
   - Clustering
   - Classification
3. Target
   - Who are the potentially bad consumers'?
   - Any common features?

OFFLINE MARKETING TACTICS ANALYSIS

1. Data
   - Media Spend
   - Prices
   - Sales
2. Key technical points
   - Marketing mix modelling
3. Target
   - The contributions of each media channel
   - ROAS(return on ad spend)

Task1: Online potentially bad consumers identification



Task1: Online potentially bad consumers identification

Data exploration
- descriptive analysis
- data type
- missing values
- outliers

Data Preprocessing
- impute
- feature engineering

Clustering
- find a proper K
- choose clustering method
- label the observation

Date splitting
- sampling

Classification
- choose classifiers
- tune parameters
- testing
- evaluation

Persistence

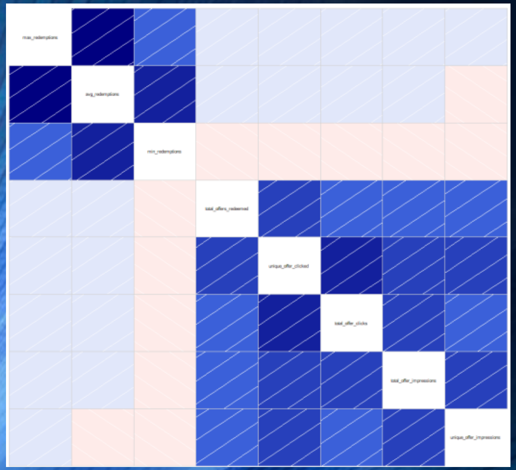
Predication

Task1: Online potentially bad consumers identification

1 Data exploration

- There are 10k records with 20 variables in the data set;
- Two of the 20 variables are categorical while others are numeric;
- The role of consumer_id is the observation identifier rather than predictors;
- Gender and customer_age have missing valuse: 4336 missing at the same time;
- Strong correlated(eg. Redemption Activity);
- Possible outliers: max(age)=119 and 169 observations by LocalOutlierFactor.

consumer_id	0
gender	4522
has_gender	0
has_first_name	0
has_last_name	0
has_email	0
has_dob	0
customer_age	5936
account_age	0
account_last_updated	0
account_status	0
app_downloads	0
unique_offer_clicked	0
total_offer_clicks	0
unique_offer_impressions	0
total_offer_impressions	0
avg_redemptions	0
min_redemptions	0
max_redemptions	0
total_offers_redeemed	0



Task1: Online potentially bad consumers identification

2 Data Preprocessing

Imputation

Gender: Missing values are treated as a separate category by itself
customer_age : KNN impute

Feature engineering

Generate "has_customer_age"
Remove customer_age and gender

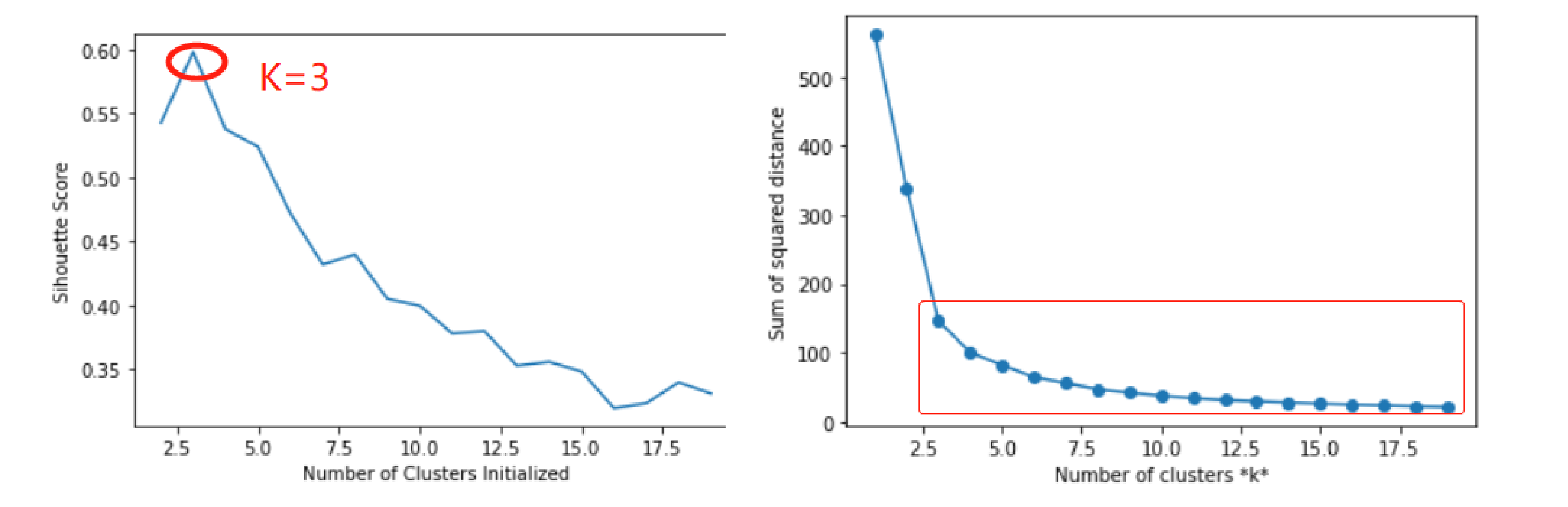
No
significant
impact

Task1: Online potentially bad consumers identification

3 Clustering

Normalization
y=(x-min)/(max-min)

Find a proper K
Silhouette_score
Elbow method



Task1: Online potentially bad consumers identification

3 Clustering

Method comparison

Methed	Silhouette score	Running time(s)
KMeans	0.5978239827225968	1.750345230102539
MiniBatchKMeans	0.5957584814547853	1.6505839824676514
hierarchical(ward)	0.5160719133161198	5.589056015014648

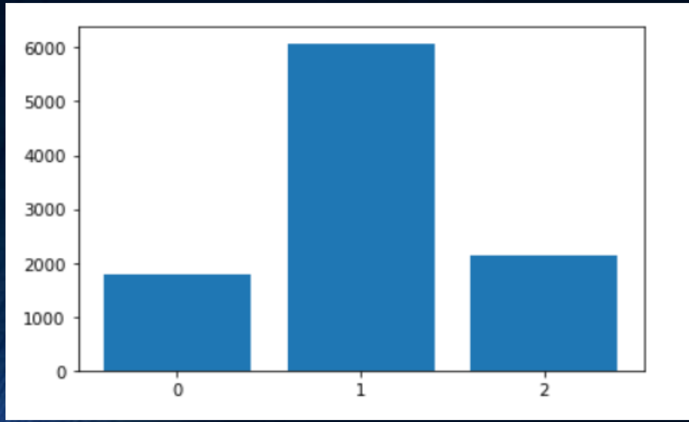
10K records sampled from a much larger training set:

MiniBatchKMeans **converges faster and** in practice this difference in quality can be quite small!

Task1: Online potentially bad consumers identification

4 Date splitting

Data Imbalance



Resampling

	random splitting	down sampling	over sampling
precision	0.9928	0.9853	0.9955
recall	0.9912	0.9904	0.9948
accuracy	0.9943	0.9910	0.9963
f1_score	0.9920	0.9878	0.9951

increases the likelihood of overfitting

Task1: Online potentially bad consumers identification

5 Multi classification

- Tuning the hyper-parameters: *GridSearchCV*

- Evaluating estimator performance: *Cross_val_score*

Methed	F1_macro	Running time(s)
Logistic regression	0.9908192778441721	0.49065351486206055
Decision tree	0.992191878064947	0.13068270683288574
RandomForest	0.9942184305075944	6.87960147857666
KNeighborsClassifier	0.9412861825983982	0.30817532539367676

LR multi_class='multinomial'

Task1: Online potentially bad consumers identification

6 Model persistence and predication

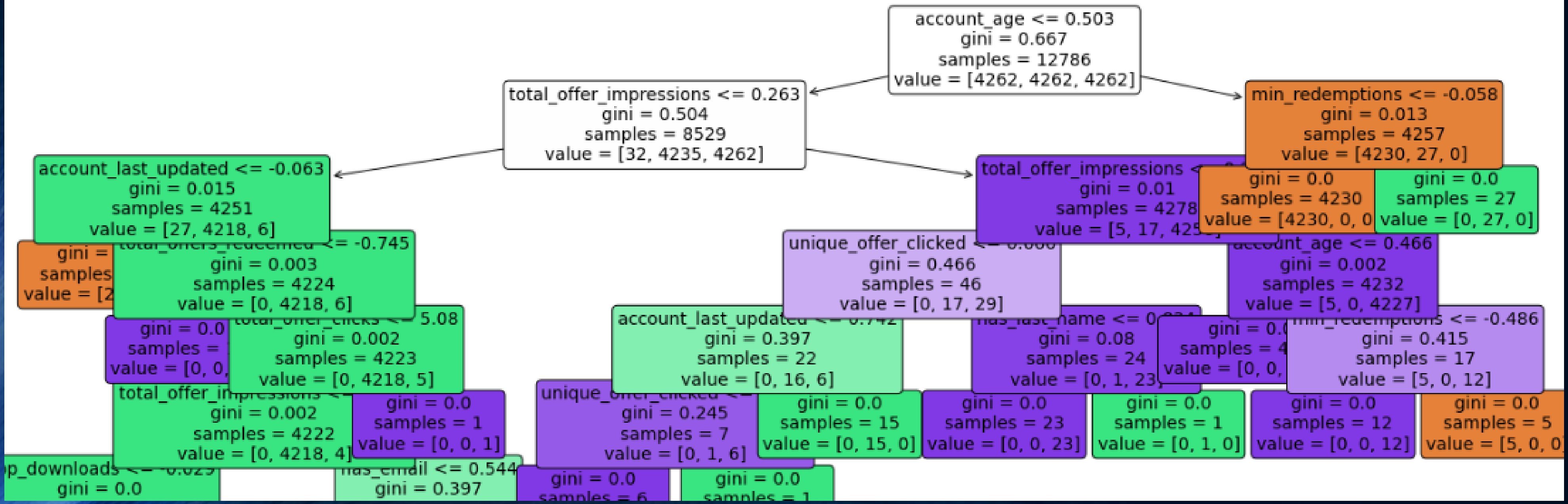
Task1: Online potentially bad consumers identification

7 Results and findings

Important feature for consumer classification(*treeModel.feature_importances_*)

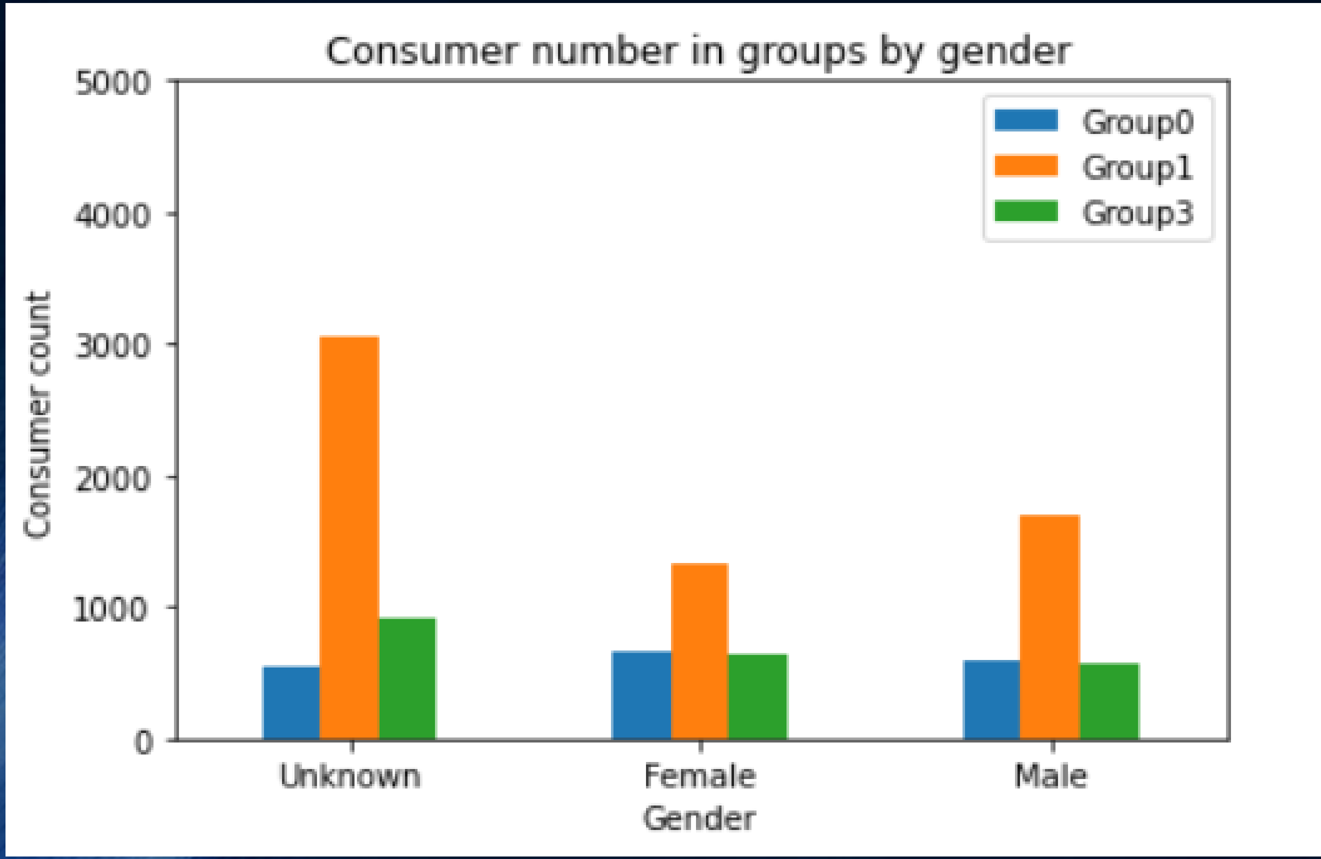
- account_age
- total_offer_impressions
- account_last_updated
- min_redemptions

Group	Average of account_age	Average of total_offer_impressions
Group0	498.279	87.829
Group1	186.318	33.591
Group2	203.343	151.549



Task1: Online potentially bad consumers identification

7 Results and findings

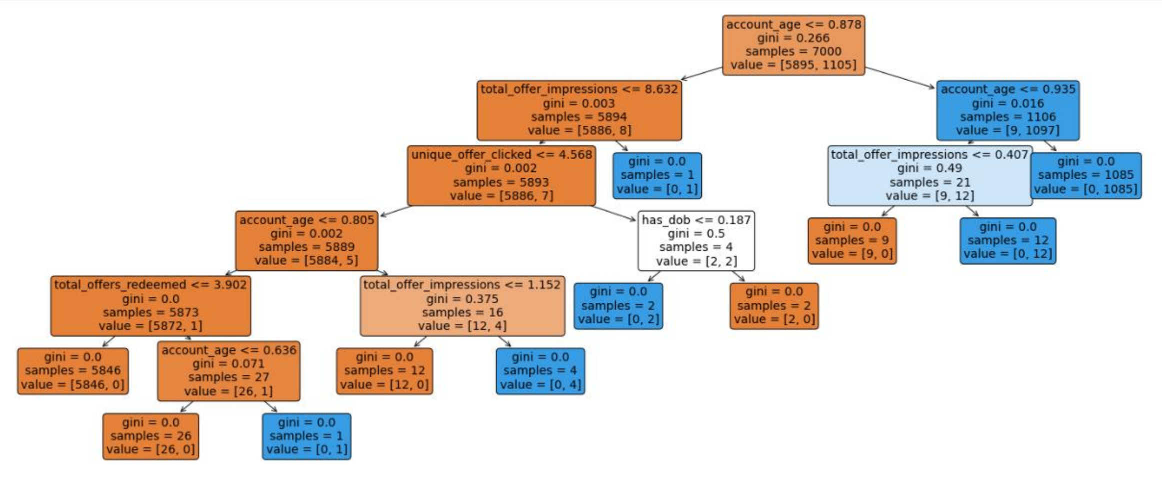


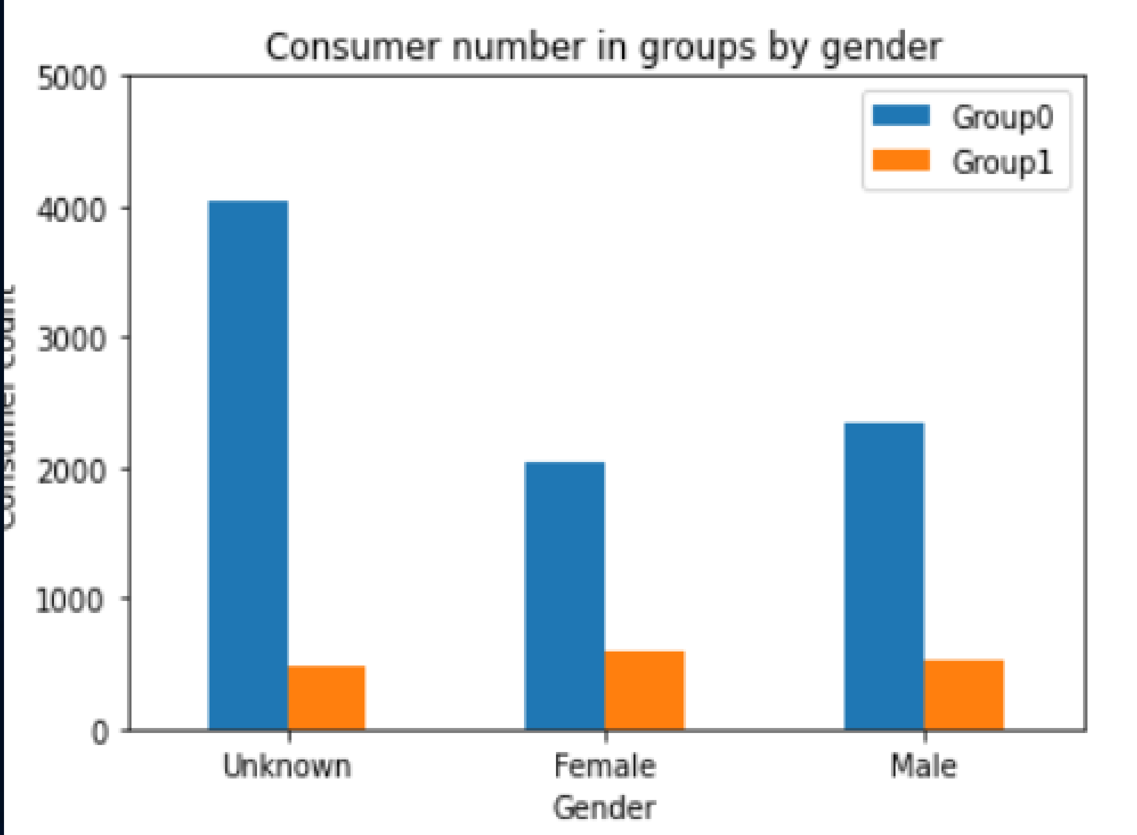
"The consumer who does not provide gender is more likely bad"

Task1: Online potentially bad consumers identification

8 other trials(K=2)

Methed	F1_macro	Running time(s)
Logistic regression	0.9980912387860279	0.17926621437072754
Decision tree	0.9943713640837519	0.07277703285217285
RandomForest	0.9943288413266078	3.7310240268707275
SGDClassifier	0.991716901548244	0.0907585620880127
LinearSVC	0.994963890271382	0.11668801307678223





Task1: Online potentially bad consumers identification

9 Possible further tasks

1. try more clustering and classification algorithms;
2. recheck the whole processing with domain knowledge;
3. consider feature selection if necessary because strong correlations;
4. monitor the model's performance (eg. ControlCharts) and update it continuously.

Task2: Offline marketing tactics analysis

Data exploration
- descriptive analysis
- data type
- missing values
- outliers
- correlation
- visualization

Data preparation
- impute
- data type
- check and transform sales

Create TimeSlices
- data splitting

Model training and tune parameters
- Adstock transform
- tune Adstock rate
- evaluation and testing
- try different models
- regression diagnostics

Results analysis
- Adstock rate
- decay effect
- contributions
- return on ad spend

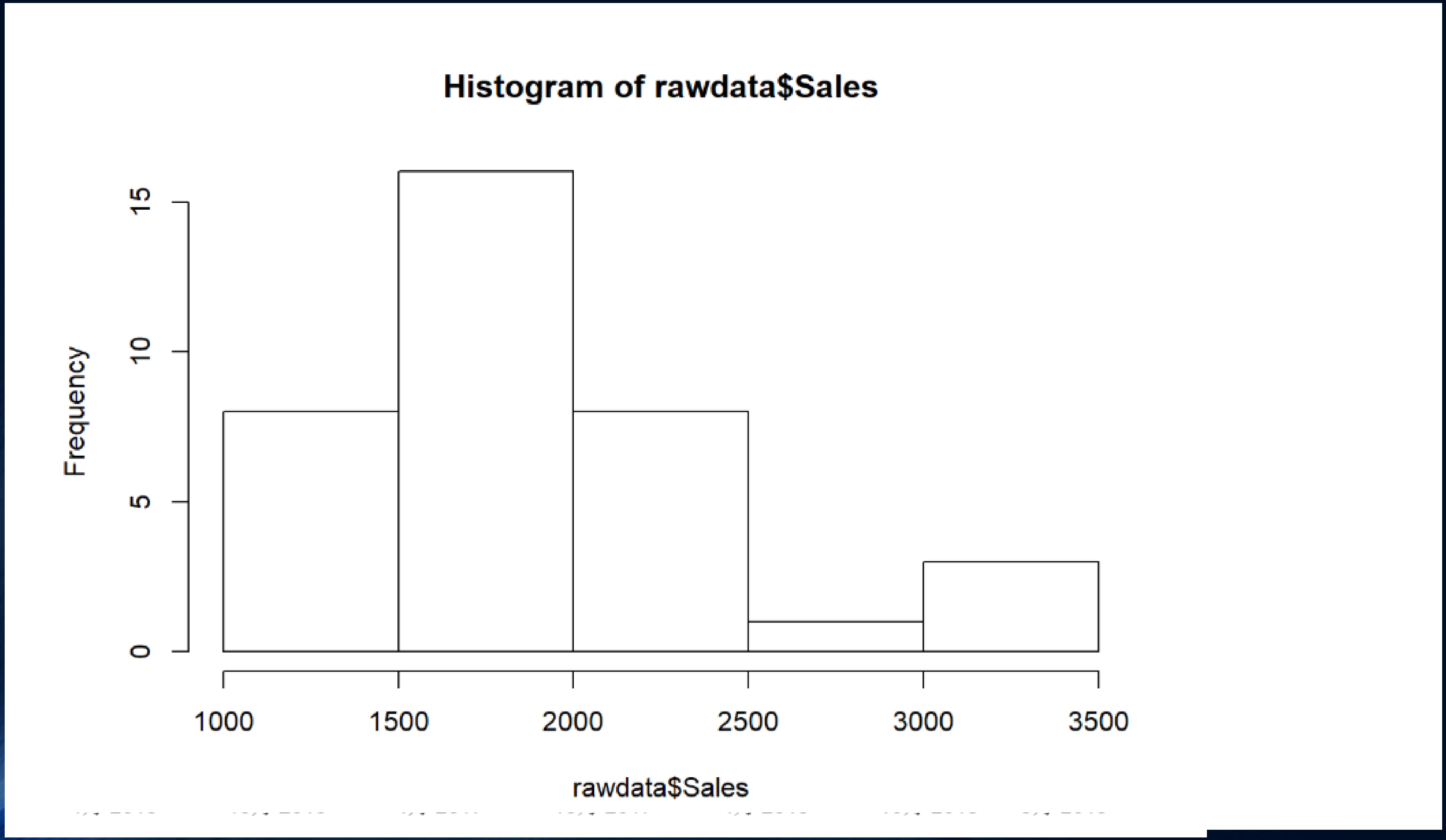
Task2: Offline marketing tactics analysis

1 Data exploration

- There are 36 records(36months) with 13 variables in time series;
- most the variables are numeric except Month;
- Price1 is constant and Price2 and Price3 changed only once during the given period;
- Months are continuous without breakpoints;
- There is a missing value in social(when month=2018/8/1).;
- Some dimensions(Pirce2&Price3,Radio&TV) are highly correlated(absolute correlation coefficient of >0.7)
- The sales shows some kind of periodicity and the right-skewed distribution means its mode is larger than mean.

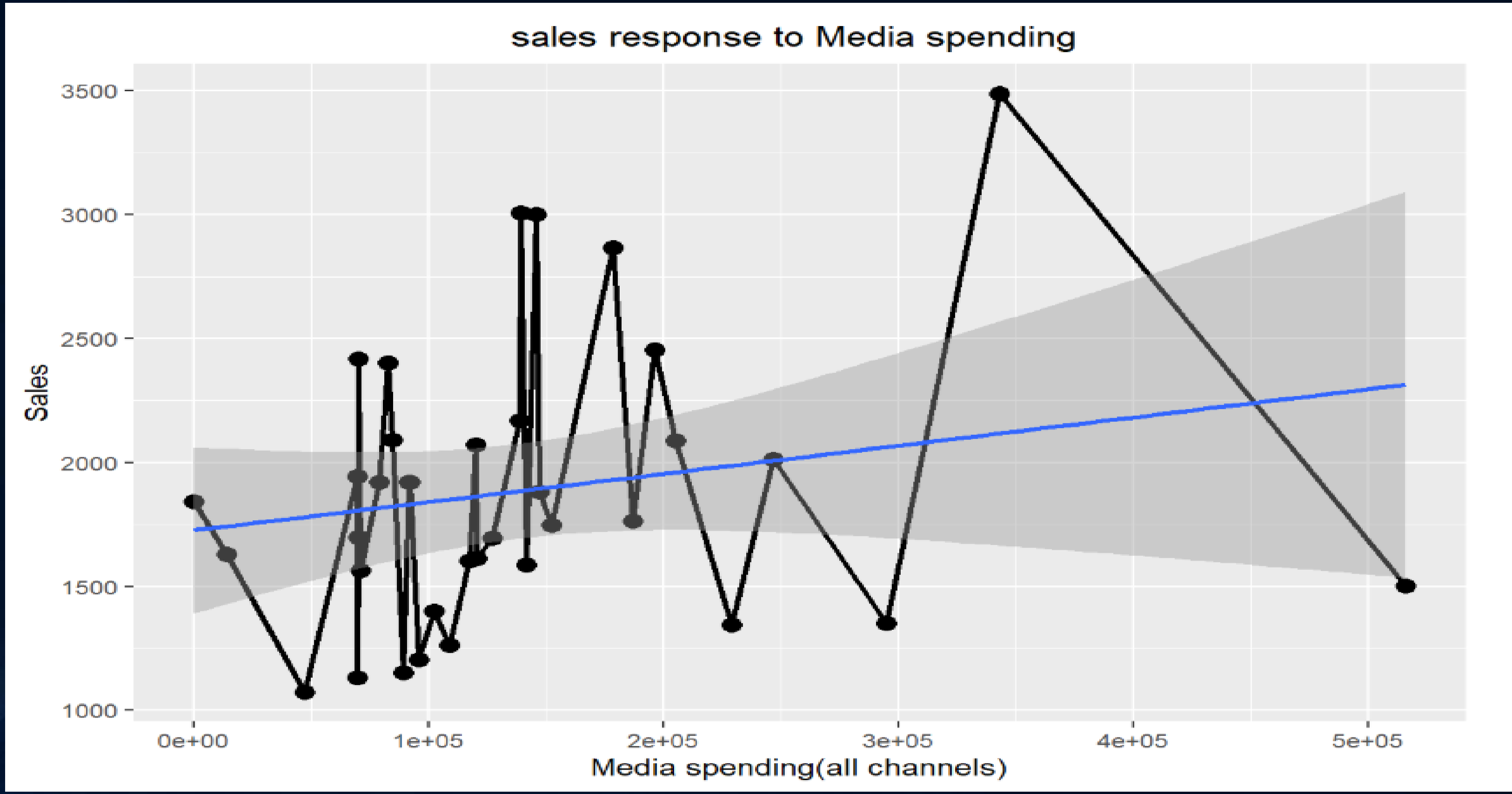
Task2: Offline marketing tactics analysis

1 Data exploration



Task2: Offline marketing tactics analysis

1 Data exploration



Task2: Offline marketing tactics analysis

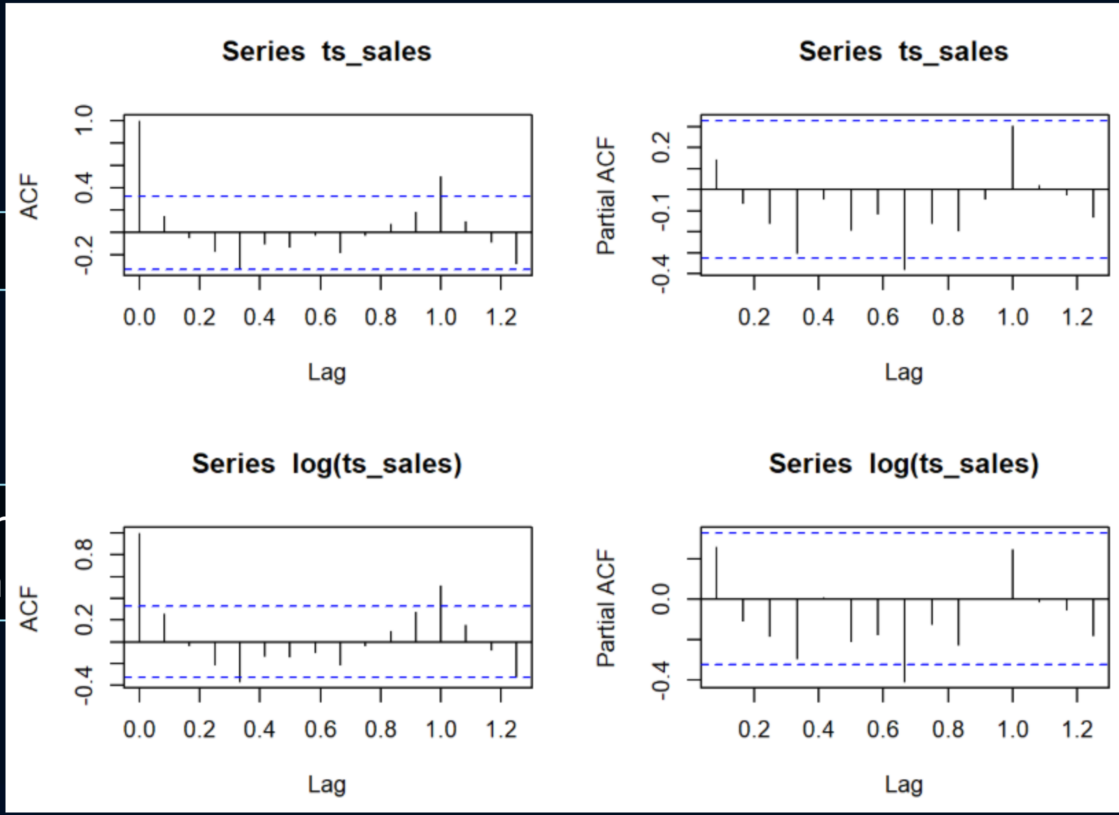
2 Data preparation

deal with missing value

Social： replace NA with 0

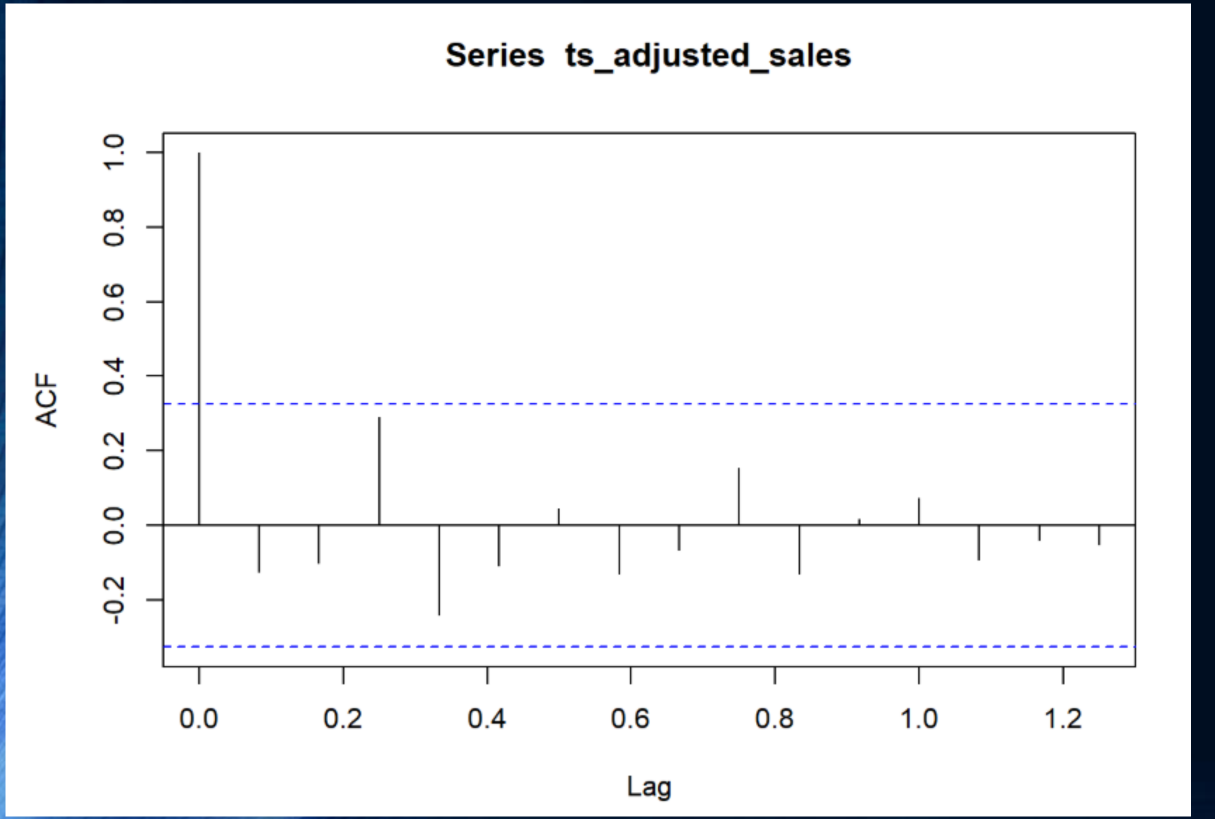
check and transform sales

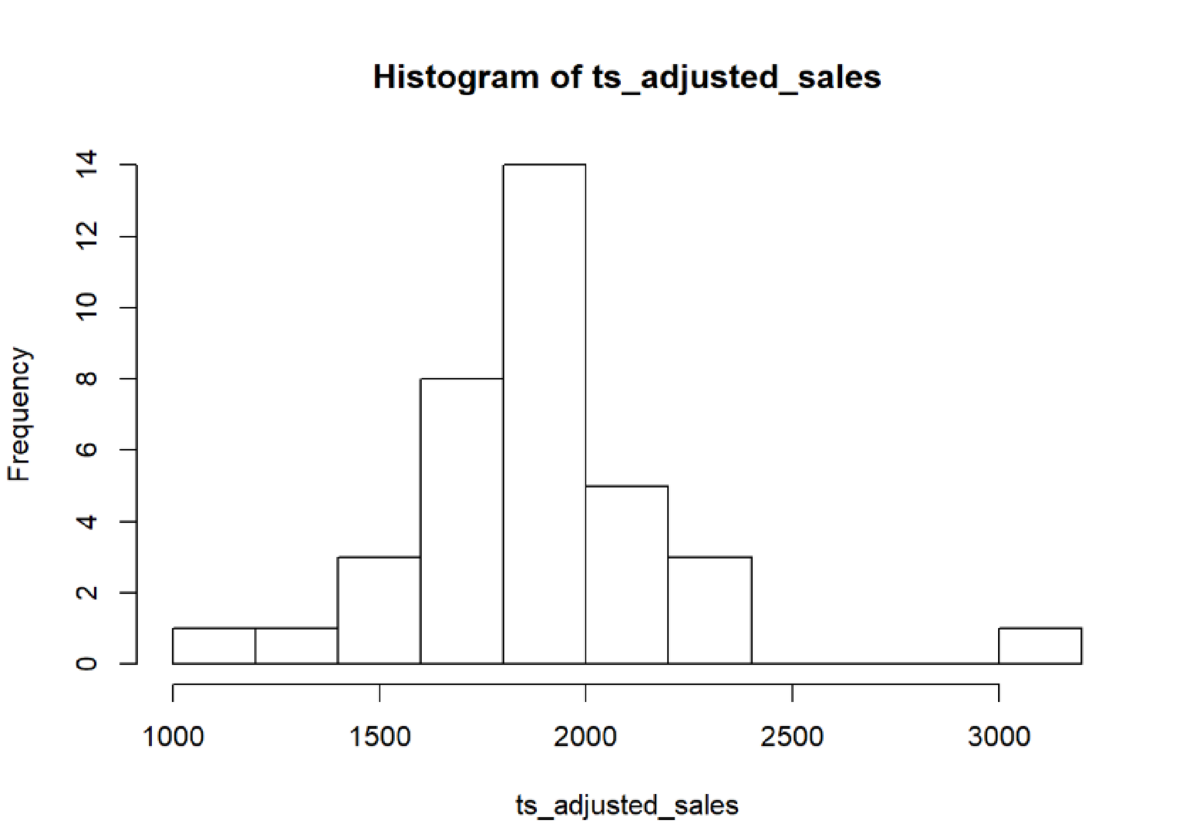
autocorrelation exits and and log transformation doesn't wor
try to decompose sales to ensure its stationarity and indepen



Task2: Offline marketing tactics analysis

2 Data preparation





ts_adjusted_sales <- ts_sales - ts_sales_components$seasonal

Task2: Offline marketing tactics analysis

3 Train and test split

- Observations in the time series are dependent:
  => the past affects the future, but the future does not affect;

- createTimeSlices(): 24 obervations for training, 6 for validation and 6 for testing .

Task2: Offline marketing tactics analysis

4 Model training and tune parameters

- Simple Decay-Effect Model

At = Tt + At-1 t=1,…., n

At is the Adstock at time t,
Tt is the value of the advertising variable at time t and is the ‘decay’ or lag weight parameter.

- Logistic (S-Curve) Decay Model

At = 1/ (1+e (-v Tt)) + At-1

the parameter v can be used to model different saturation levels.

Adstock transform:

Task2: Offline marketing tactics analysis

4 Model training and tune parameters

- adstock_Rates=seq(from=0, to=1, by=0.05)
- each advertising variable has it's own adstock_Rate or the same
- nested Cross-Validation(7 loops)

Tune Adstock rate

Task2: Offline marketing tactics analysis

4 Model training and tune parameters

The best model:

- each advertising variable has it's own adstock_Rate
- using Simple Decay-Effect transformation
- $R2 = 0.6035$, so 60.36% of the variability of the response data can bee explained by the model.

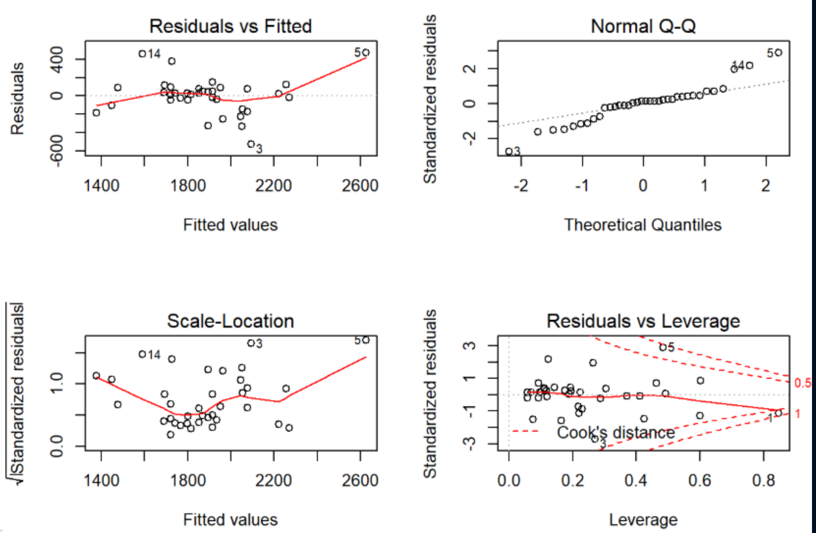
best model:Criterion:R2

Task2: Offline marketing tactics analysis

4 Model training and tune parameters

```
#modeling
media_ad=c(Magazine.adstock,Newspaper.adstock,Radio.adstock,OOH.adstock,TV.adstock,Search.adstock,Display.adstock,Social.ads
tock)
modFit.4 <-lm(adjusted_sales~.,data=media_ad)
summary(modFit.4)
```

```
##
## Call:
## lm(formula = adjusted_sales ~ ., data = media_ad)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -525.93  -59.04   27.34   80.98  470.71
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)   1.546e+03  2.271e+02   6.806 2.61e-07 ***
## X_ad.Magazine  2.047e-03  1.565e-03   1.308 0.201913
## X_ad.Newspaper 3.521e-03  3.744e-03   0.941 0.355265
## X_ad.Radio     1.099e-02  2.924e-03   3.759 0.000835 ***
## X_ad.OOH      -5.569e-05  9.891e-04  -0.056 0.955511
## X_ad.TV       -1.403e-02  6.009e-03  -2.334 0.027256 *
## X_ad.Search    1.979e-04  2.363e-03   0.084 0.933868
## X_ad.Display  -1.192e-02  3.298e-03  -3.614 0.001216 **
## X_ad.Social    2.035e-03  1.742e-03   1.168 0.252954
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 226.2 on 27 degrees of freedom
## Multiple R-squared:  0.6036, Adjusted R-squared:  0.4862
## F-statistic:  5.14 on 8 and 27 DF,  p-value: 0.0005854
```



Task2: Offline marketing tactics analysis

5 Model testing

```
##       R2_avg RMSE_avg  MAPE_avg
## 1 0.6426396 272.7579 0.1201196
```

Task2: Offline marketing tactics analysis

6 Results analysis

Adstock Affect: the prolonged or lagged effect of advertising on consumer purchase behavior.

There are two dimensions to advertising adstock:

- Decay effect:the impact of past advertisement on present sales;
- saturation or diminishing returns effect.

```
##       ChannelName AdstockRate
## [1,] "Magazine"   "0.6"
## [2,] "Newspaper"  "1"
## [3,] "Radio"      "0"
## [4,] "OOH"        "1"
## [5,] "TV"         "0"
## [6,] "Search"     "0.6"
## [7,] "Display"    "0"
## [8,] "Social"     "0.35"
```

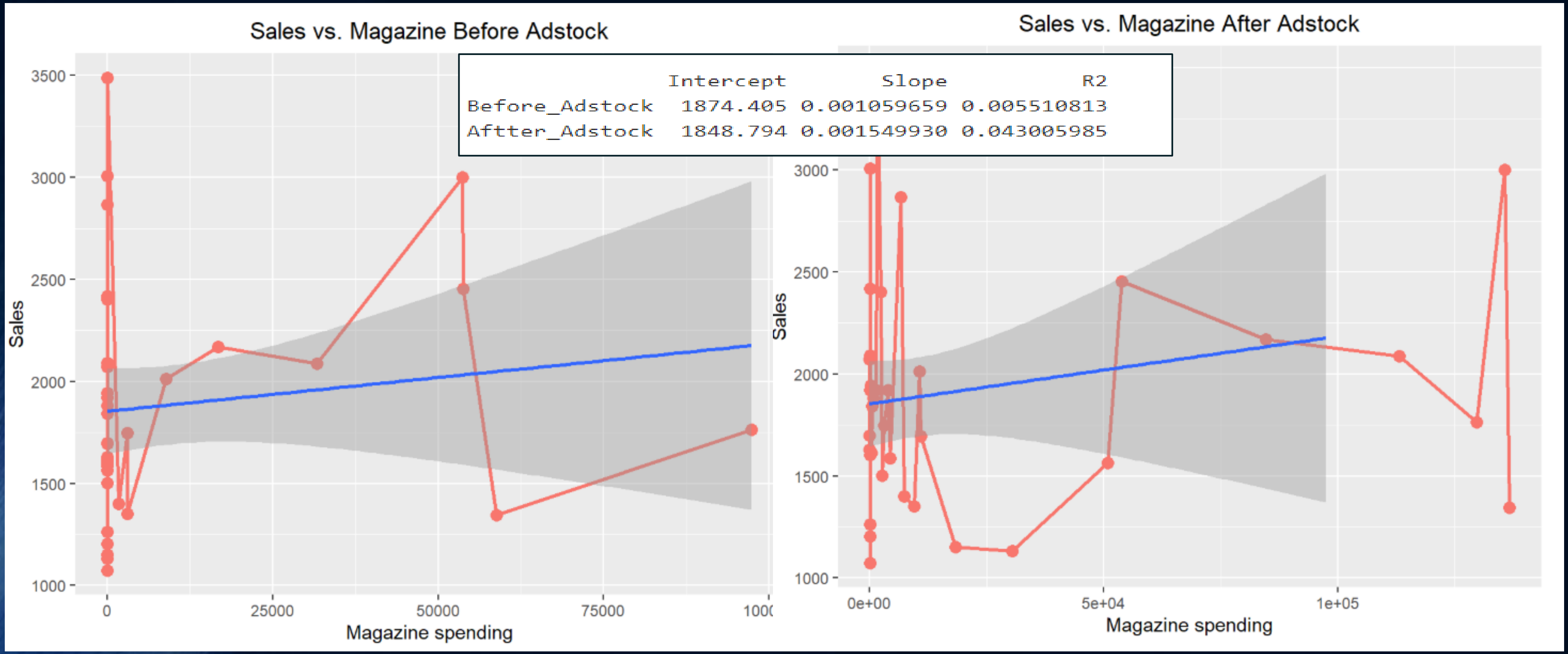
Magazine;
Newspaper;
OOH;
Search;
Social.

Decay effect

Task2: Offline marketing tactics analysis

6 Results analysis

Take the magazine as an example,

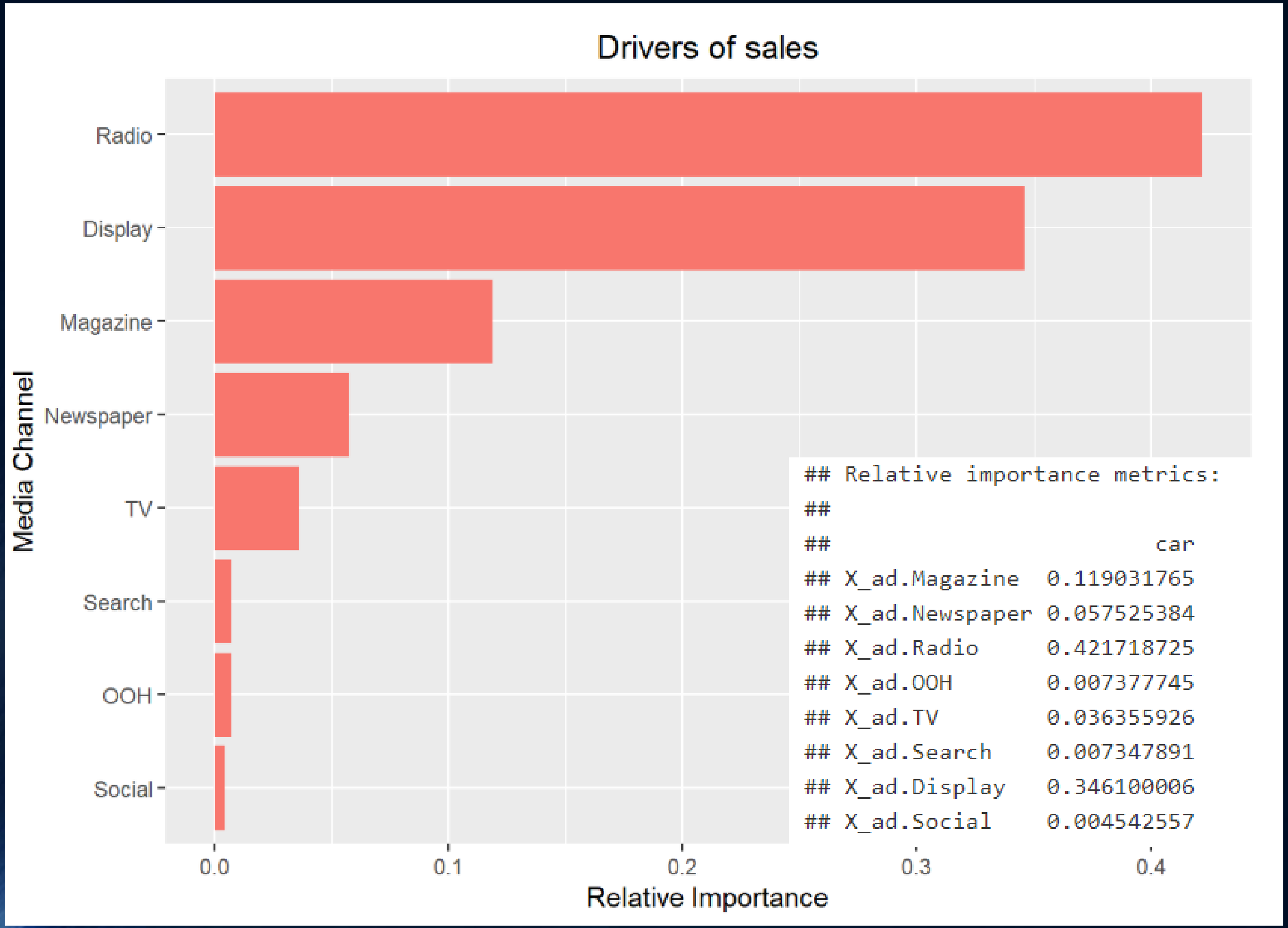


	Intercept	Slope	R2
Before_Adstock	1874.405	0.001059659	0.005510813
Aftter_Adstock	1848.794	0.001549930	0.043005985

Decay effect

Task2: Offline marketing tactics analysis

6 Results analysis



Contribution Charts

Task2: Offline marketing tactics analysis

6 Results analysis

$$\frac{\text{Revenue Generated by Ads}}{\text{Cost of Ads}} = \text{ROAS}$$

ROAS: based on sales

```
##   X_ad.Magazine X_ad.Newspaper     X_ad.Radio        X_ad.OOH         X_ad.TV
##    0.0245531524   0.0358428538   0.0390312961    0.0011123616    0.0143801338
##     X_ad.Search    X_ad.Display    X_ad.Social
##    0.0003828049   0.0311106194   0.0002703955
```

ROAS: Sales*(Price1+Price2+Price3)

```
##   X_ad.Magazine X_ad.Newspaper     X_ad.Radio        X_ad.OOH         X_ad.TV
##      3.93732533     5.74773349     6.25902975      0.17837748      2.30598761
##     X_ad.Search    X_ad.Display    X_ad.Social
##      0.06138631     4.98887590     0.04336042
```

is the unit of sales the same as media spend?

return on ad spend

Task1: Online potentially bad consumers identification

7 Questions and further reach

- Considering the prices;
- How about feature selection;
- Any other optimization;
- Is the ROAS calculation correct?
- Why negative slope?

Thanks!