Group 3

Introduction to Data Science

Instructors:
Kiều Vũ Minh Đức - Lê Nhựt Nam

Data Science

Members

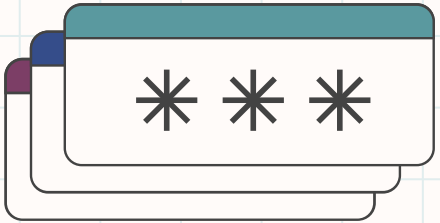
ID Student	Full Name
20127458	Đặng Tiến Đạt
20127627	Nguyễn Quốc Thắng
20127680	Phạm Thị Ánh Phát

CONTENT

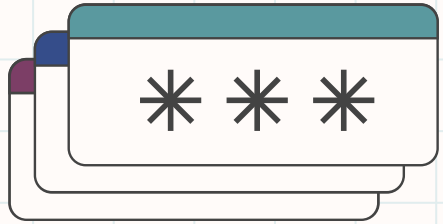
Group 3

01

About project



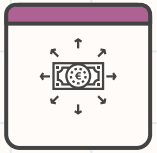
About project



- Data is crawled for analysis, statistics and sales to help Tiki’s sellers supply and demand adjustment.

- Predict the number of products sold in Tiki.

About Data



p_id	Id of the product
p_name	Name of the product
p_id_shop	Shop id that sells the product
p_shop_name	Shop name that sells the product
p_brand	Brand of the product
p_categories	Category of product

p_day_create	Number of days the product was created since data collection
p_sold_quantity	Number of the product sold
p_original_price	Original price of the product
p_current_price	Current price of the product
p_discount_rate	Discount rate of the product

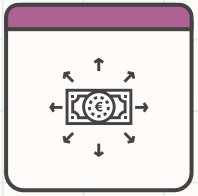
Group 3

02

Questions

We will create questions and derive insight the data.

Meaningful Questions

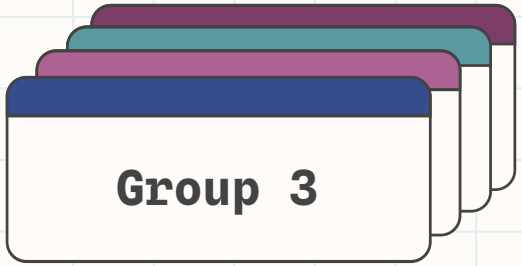


Question 1:

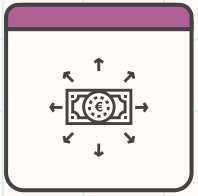
Which store sells best and what items in that store sell best? Why is it the best seller?

Question 2:

Sales of each category in “Tiki trading”. The best-selling category accounts for what percentage of that sales?

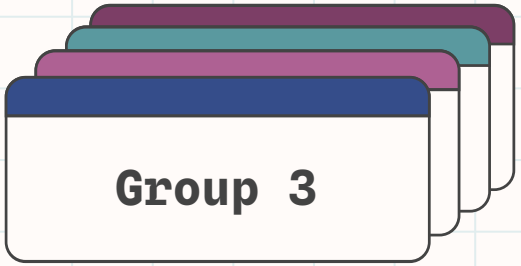


Meaningful Questions



Question 3:

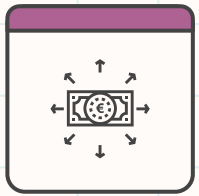
Compared with the discount and no discount, which shop's items sell the best?



Question 4:

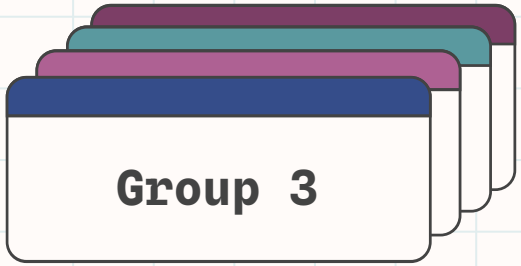
Calculate the average daily sales of the brand. Bring out the brand with the highest and lowest revenue.

Meaningful Questions



Question 5:

What products do brands have? Compare sales of the same product but in different brands.

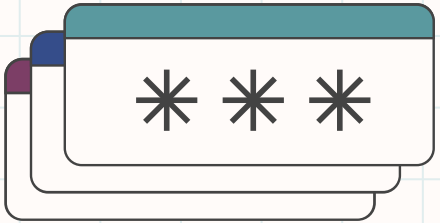


Group 3

03

Model & Evaluation

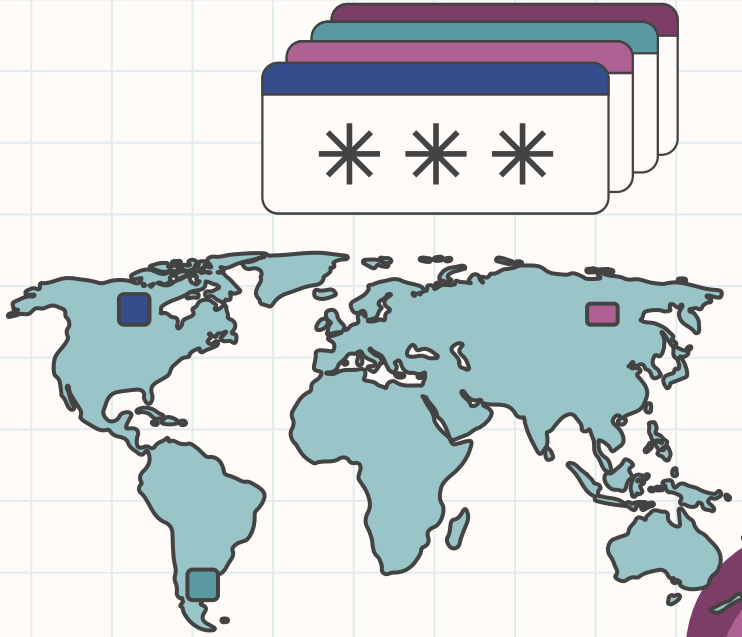
Use model to predict the number of products sold in Tiki.



Models

Linear Regression

Polynomial Regression



Linear Regression

```
# Create a pipeline
pipe = Pipeline([('scaler', StandardScaler()), ('linear', LinearRegression())])
param_grid = {'linear__fit_intercept': [True, False], 'linear__normalize': [True, False]}
```

Features

- Use all features in canonical data

Parameters

- `Fit_intercept: False`
- `Normalize: True`

❏ **Comments:**

- RMSE and MSE in Linear Regression depend on test rate. When it has proper test rate, result is improved.

Result with hyperparameters

`Test_rate: 0.2`

```
MSE: 1.1889984435678778
MAE: 0.4594519653958411
RMSE: 1.0904120521930587
```

`Test_rate: 0.3`

```
MSE: 0.8979550530061658
MAE: 0.43993341090654964
RMSE: 0.9476049034308369
```

Linear Regression

```
# Create a pipeline
pipe = Pipeline([('scaler', StandardScaler()), ('linear', LinearRegression())])
param_grid = {'linear__fit_intercept': [True, False], 'linear__normalize': [True, False]}
```

Features

- Drop brand and categories

Parameters

- `Fit_intercept: False`
- `Normalize: True`

Result with hyperparameters

`Test_rate: 0.2`

```
MSE: 1.2085987639621993
MAE: 0.5122997486453387
RMSE: 1.0993628900241263
```

`Test_rate: 0.3`

```
MSE: 0.9149633139651127
MAE: 0.4912052813986782
RMSE: 0.9565371471956082
```

❏ **Comments:**

- RMSE and MSE in Linear Regression is really high because the features in data are not linear with each other.(as figure in code)

Polynomial Regression

```
param_grid = {
    'polynomialfeatures__degree': [3,4,5],
    'linearregression__fit_intercept': [True, False],
    'linearregression__normalize': [True, False]
}
```

Features

- Use all features in canonical data

Parameters

- Degree: 3
- Fit_intercept: False
- Normalize: True

Result with hyperparameters

Test_rate: 0.2

```
MSE:  0.6021217736471877
MAE:  0.3555188131847011
RMSE:  0.7759650595530624
```

Test_rate: 0.3

```
MSE:  0.5976363481591647
MAE:  0.3782701648986338
RMSE:  0.7730694329484026
```

Polynomial Regression

```
param_grid = {
    'polynomialfeatures__degree': [3,4,5],
    'linearregression__fit_intercept': [True, False],
    'linearregression__normalize': [True, False]
}
```

Features

- Drop brand and categories

Parameters

- Degree: 3
- Fit_intercept: False
- Normalize: True

❏ **Comments:**

- RMSE and MSE in Polynomial Regression is still high but better than Linear Regression.

Result with hyperparameters

Test_rate: 0.2

```
MSE:  0.5952179757076738
MAE:  0.3451881277197315
RMSE:  0.7715037107543125
```

Test_rate: 0.3

```
MSE:  0.5452448773234302
MAE:  0.3585357713929272
RMSE:  0.7384069862368788
```

Group 3

04

Reflection

For each member re-evaluate the work

If we have more time...

We will

- Create more questions and derive important, useful insights from the data.
- Use more techniques to find hyperparameters
  - Such as Bayes search...
- Predict top trending or not
  - May be using Logistic Regression
- Compare with more model to find the optimal result
  - Such as Bayesian Linear Regression,...

For each member…

ID Student: 20127458 - Đặng Tiến Đạt

What difficulties have you encountered?

- **Github**: merge code with notebook files (conflict,...)
- **For each step:**
  - **Data engineer:** Data exchange protocols, choose the right API for you and the params, headers, ... required to be able to request the API. Data cleaning is not good, leading to errors when the model
  - **Data analyst:** Choose how to visualize to be reasonable for the question you posed, statistical methods to find that question.
  - **Machine learning:** Choose the right algorithm models to make accurate predictions, scale the data as well as create a pipeline to process each different thread, how to evaluate the model, which model will be selected when completed.

What have you learned?

- Understand the workflow of a data science as well as be more careful in each step so that the next steps do not affect much
- Proficient in using project libraries (numpy, pandas, matplotlib, ...)

For each member…

ID Student: 20127627

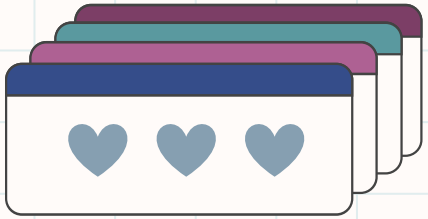
What difficulties have you encountered?

- About data collection, I had difficulty in crawling data by getting API.Finding data and how to crawl using the API is also the first time, so it is quite difficult.
- About clear and preprocess data: It is quite difficult to clear data from crawl data, data is missing, data type is wrong a lot, making it quite difficult for me to keep or delete that row. And about the examination as well as asking questions and answering, as this is the first time, it is a challenge to ask meaningful questions
- About the model, because I have not studied machine learning, it is quite difficult to choose a model and use it as input, and the model's prediction is wrong, the difference is too much, it is also unpredictable.

What have you learned?

- Crawl data using API
- Learn how to ask questions based on data and visualize questions correctly
- Steps to clean data and explore data
- Learn how to model data, based on the given data to try to predict a feature

For each member…



ID Student: 20127680 - Phạm Thị Ánh Phát

What difficulties have you encountered?

- Because Tiki does not provide API so we have manually crawl data for each page.
- Because i had not learned model before, so i have difficult in using model to predict. But now, i can :D

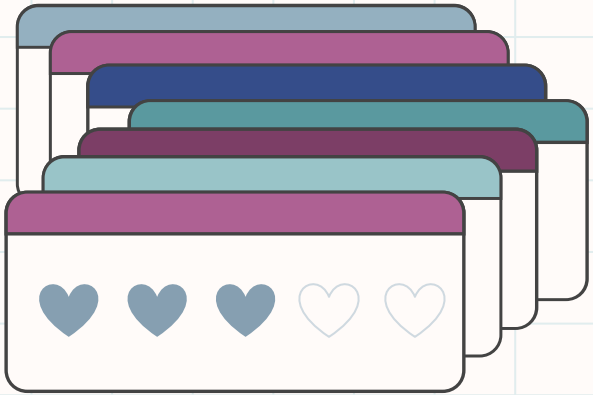
What have you learned?

- Learn the process of Data Analysis. From raw data to visualize data for customers.
- Learn how to analysis information of products in order to bring benefits in business.
- Learn how to use pipeline in preprocessing.
- Learn how to use techniques to evaluate predicted models.

Group 3

Thanks!

Do you have any questions?



References

- https://machinelearningcoban.com/2016/12/28/linearregression/
- https://scikit-learn.org/stable/modules/generated/sklearn.pipeline.Pipeline.html
- Two files demo assignment in class.
- https://slidesgo.com/theme/pestel-analysis-thesis-defense?fbclid=IwAR1Ju4MeCCwSeQ-v7PPUJEE0do0IPxpMTB4FVhLm3zmzq0AEx8GWqd1ptks#position-132&results-11706
- https://www.kaggle.com/datasets/hellbuoy/car-price-prediction