Online Shoppers Purchasing Intention Classification

Team 13:
Guangze Wang
Jackie Du
Ziyuan Wu
Miao Fang

TABLE OF CONTENTS

01
Introduction

Executive Summary

PROBLEM

E-commerce website managements hope to identify online customers with higher purchasing intentions by analyzing their visiting behaviors on the website, so that they could tailor their promotions, website design, etc accordingly to improve ROI.

PROPOSED SOLUTION

We built 4 machine learning models to classify online shoppers:

- Logistic Regression
- Naive Bayes
- LightGBM
- Multilayer Perceptron

Accuracy is used to evaluate the model performance.

CONCLUSION & NEXT STEP

LightGBM is the best performing model.

Model performance could be improved by collecting more data, further feature engineering, model tuning, etc.

Problem Statement

The increasing prevalence of e-commerce has created great potential in the market. Purchase conversion rate is a key metric measuring the success of an e-commerce website. Managements of e-commerce website would hope to analyze online shoppers' visiting behaviors on the website and predict customers' intention to finalize the transaction. With information on which group of customers having higher purchasing intention, the company can present customized promotions/campaign, further boosting purchasing behaviors and sales.

Project Objective

The purchase intention model will be a binary classification problem (1 for purchase made; 0 for no purchase made). Our implementation will be a machine learning pipeline that incorporates data cleaning, feature engineering and classification models. Baseline model will be defined, as well as improved models in terms of metrics performances and business need. The result will be a fine-tuned best model and following reports to provide customer intention predictions and corresponding business insights.

Data Summary

DATA SOURCE: UCI Machine Learning Repository

DATA INFORMATION

- E-commerce dataset that describes behaviors of online shoppers and their purchasing intention
- 12330 rows without missing values
- 18 variables (10 numerical and 8 categorical)
- Target variable: Revenue (True - purchase made/ False - no purchase made)
- Abstract landing pages with technology devices

DATA ATTRIBUTES

- Administrative & Duration
- Informational & Duration
- Product Related & Duration
- Bounce Rates
- Exit Rates
- Page Values
- Special Day
- Month
- Operating System
- Browser
- Region
- Traffic Type
- Visitor Type
- Weekend
- Revenue

Data Example

	Administrative	Administrative_Duration	Informational	Informational_Duration	ProductRelated	ProductRelated_Duration	BounceRates	ExitRates	PageValues
0	0	0.0	0	0.0	1	0.000000	0.200000	0.200000	0.000000
1	0	0.0	0	0.0	2	64.000000	0.000000	0.100000	0.000000
2	0	0.0	0	0.0	1	0.000000	0.200000	0.200000	0.000000
3	0	0.0	0	0.0	2	2.666667	0.050000	0.140000	0.000000
4	0	0.0	0	0.0	10	627.500000	0.020000	0.050000	0.000000
...	...	...	...	...	...	...	...	...	...
12325	3	145.0	0	0.0	53	1783.791667	0.007143	0.029031	12.241717
12326	0	0.0	0	0.0	5	465.750000	0.000000	0.021333	0.000000
12327	0	0.0	0	0.0	6	184.250000	0.083333	0.086667	0.000000
12328	4	75.0	0	0.0	15	346.000000	0.000000	0.021053	0.000000
12329	0	0.0	0	0.0	3	21.250000	0.000000	0.066667	0.000000

12330 rows × 18 columns

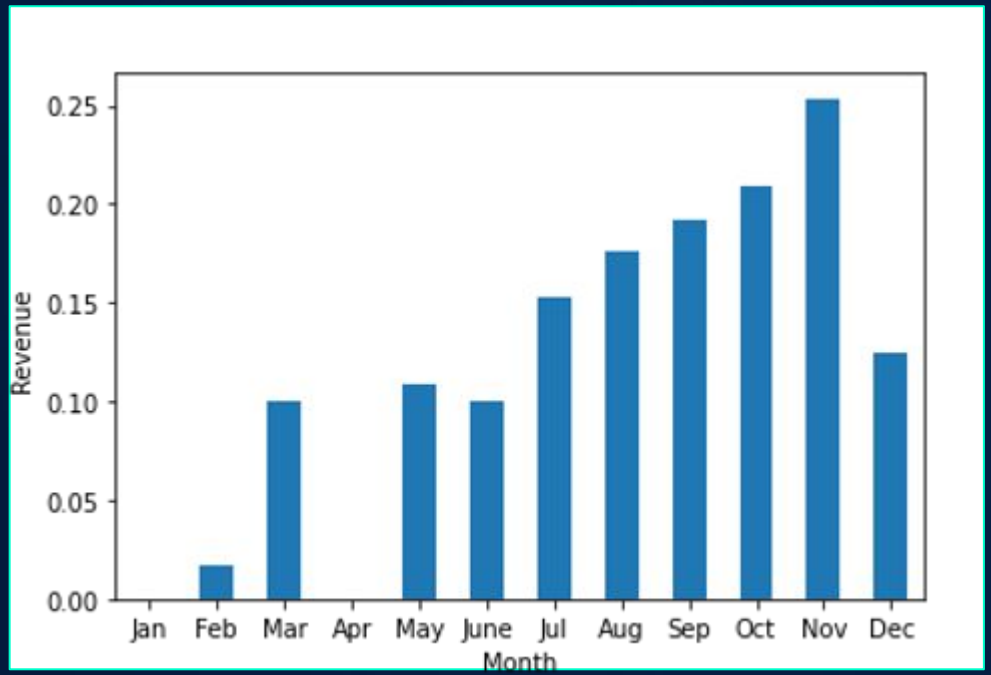
SpecialDay	Month	OperatingSystems	Browser	Region	TrafficType	VisitorType	Weekend	Revenue
0.0	Feb	1	1	1	1	Returning_Visitor	False	False
0.0	Feb	2	2	1	2	Returning_Visitor	False	False
0.0	Feb	4	1	9	3	Returning_Visitor	False	False
0.0	Feb	3	2	2	4	Returning_Visitor	False	False
0.0	Feb	3	3	1	4	Returning_Visitor	True	False
...	...	...	...	...	...	...	...	...
0.0	Dec	4	6	1	1	Returning_Visitor	True	False
0.0	Nov	3	2	1	8	Returning_Visitor	True	False
0.0	Nov	3	2	1	13	Returning_Visitor	True	False
0.0	Nov	2	2	3	11	Returning_Visitor	False	False
0.0	Nov	3	2	1	2	New_Visitor	True	False

02

Exploratory Data Analysis

EDA-1

The data is clean and has no null values. However, it doesn't have records in some months.

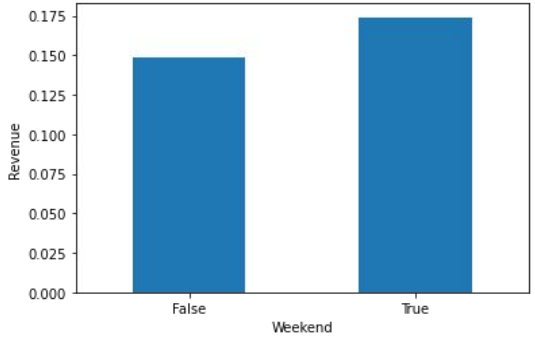



```
np.sum(df.isna())
```

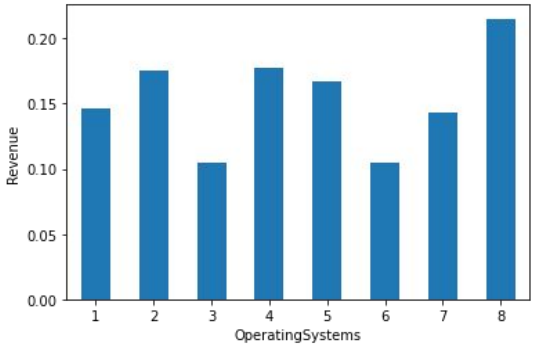
Administrative	0
Administrative_Duration	0
Informational	0
Informational_Duration	0
ProductRelated	0
ProductRelated_Duration	0
BounceRates	0
ExitRates	0
PageValues	0
SpecialDay	0
Month	0
OperatingSystems	0
Browser	0
Region	0
TrafficType	0
VisitorType	0
Weekend	0
Revenue	0

EDA-2

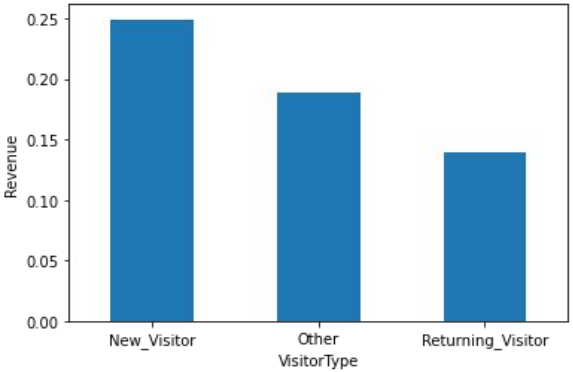
From EDA we can see that our data is distributed evenly in weekends and operating systems. However, there're missingness in months, particularly there's no data in January and April. We believe that this missingness will not affect our role in classifications.



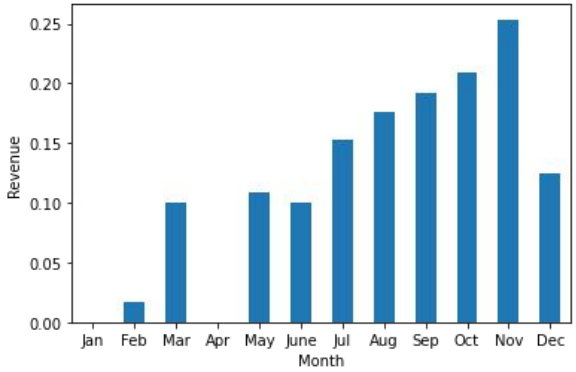
Weekend has more revenue



Operating Systems might relate to revenue



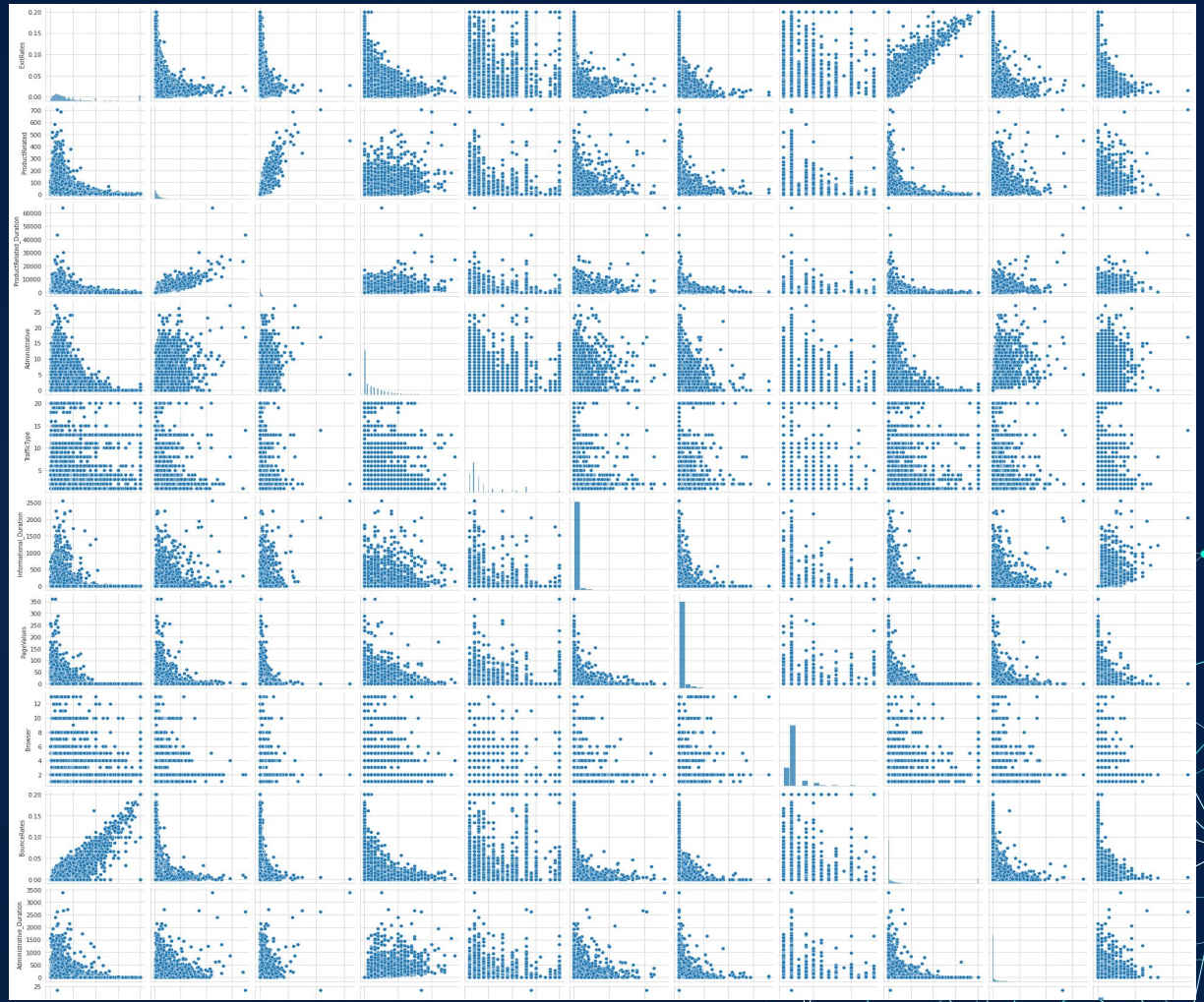
New visitor contributes more to revenue



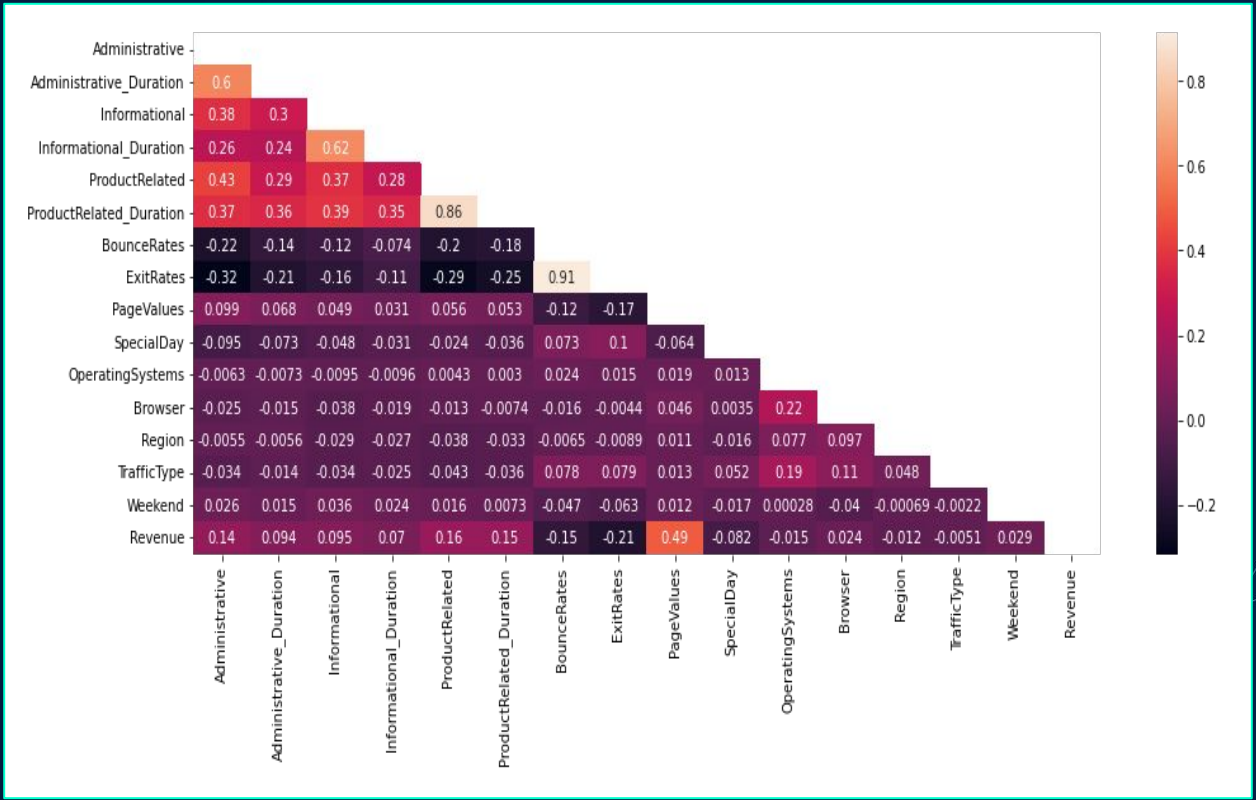
Revenue shows seasonality

EDA-3

By plotting the paired graphs, we can see there exists correlations between our features, influencing the i.i.d assumptions. Feature engineering might be needed to solve the potential multicollinearity problems.



EDA-4



We plot the correlations between variables, and set up a threshold(0.01) to drop any feature that has minimal correlations with the revenue. We assume that rest of the features can contribute to classifying revenues.

03

Insights

Assumption

- Assume columns 'OperatingSystems', 'Browser','Region','TrafficType', and 'Weekend' have limited influence on predicting if an online shopper will purchase
- Assume January and April do not have data, which reflects the real situation

Data Preparation

- Change data types for columns
- Drop features with low correlations

Consideration

Model Metrics

Accuracy

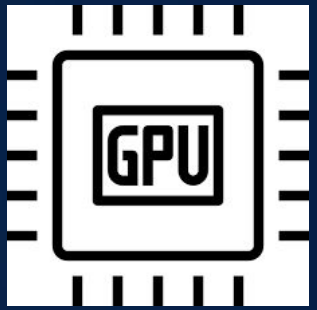
Feature Engineering

- **One Hot Encoding** for categorical variables
- **Standard scale** for numerical variables
- Perform **SMOTE** on training data to handle imbalance dataset

04

Modeling

Tooling and Hardware



Models in this project is using sklearn based on Google Colab with GPUs.

Logistic Regression

- Probabilistic classifier to predict a binary outcome
- Baseline model

```
from sklearn.linear_model import LogisticRegression

log_reg = LogisticRegression()
log_reg.fit(x_train, y_train)
```

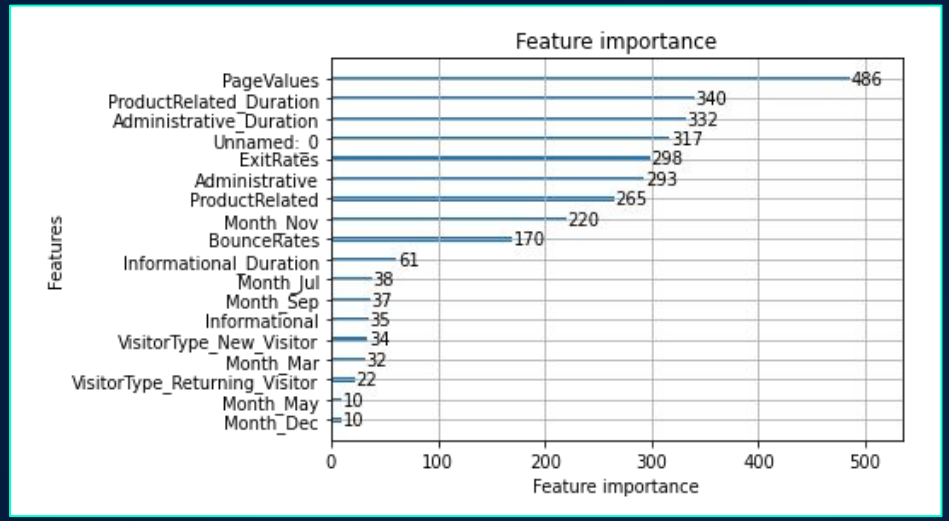
Naive Bayes

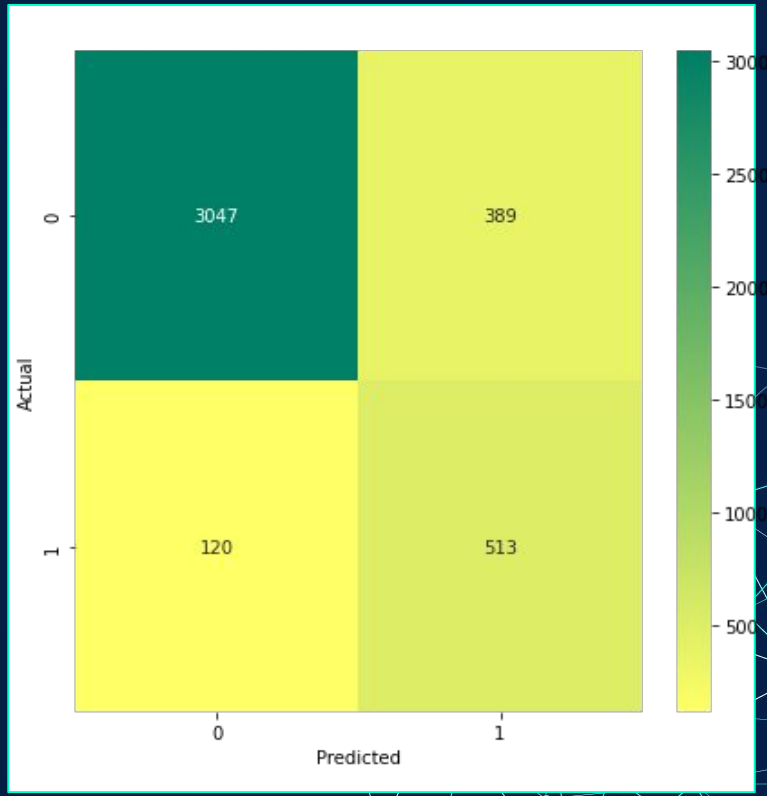
- Probabilistic classifier based on Bayes' theorem with strong (naive) independence assumptions between features

```
from sklearn.naive_bayes import BernoulliNB

nb = BernoulliNB()
nb.fit(x_train, y_train)
```

LightGBM



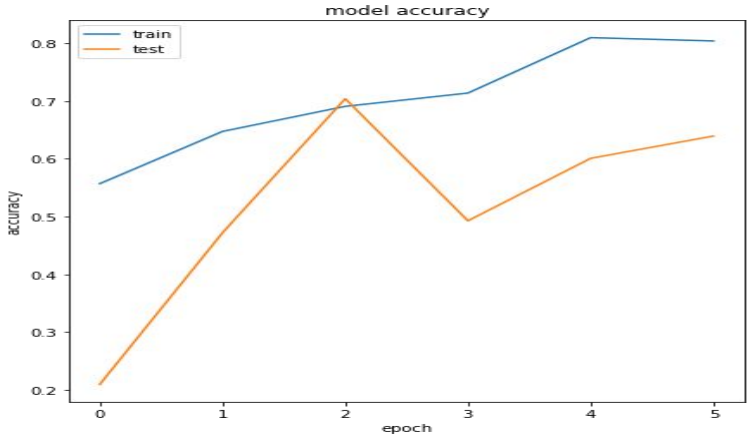


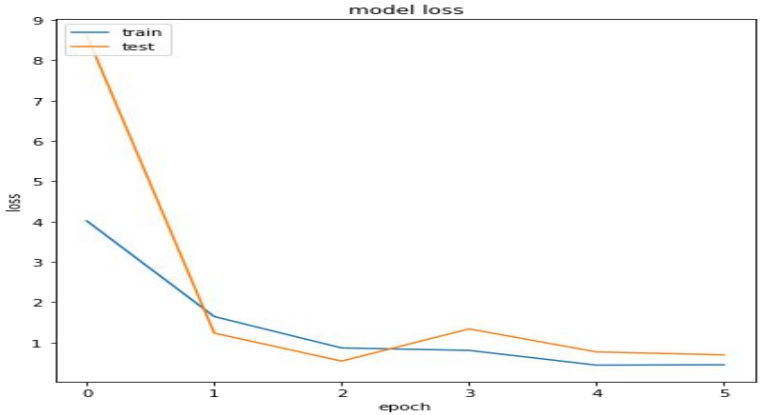
	precision	recall	f1-score	support
0	0.96	0.89	0.92	3436
1	0.57	0.81	0.67	633
accuracy			0.87	4069
macro avg	0.77	0.85	0.80	4069
weighted avg	0.90	0.87	0.88	4069

Multilayer Perceptron-1

Layer (type)	Output Shape	Param #
dense_19 (Dense)	(None, 150)	3600
dense_20 (Dense)	(None, 150)	22650
dense_21 (Dense)	(None, 150)	22650
dense_22 (Dense)	(None, 100)	15100
dense_23 (Dense)	(None, 100)	10100
dense_24 (Dense)	(None, 100)	10100
dense_25 (Dense)	(None, 50)	5050
dense_26 (Dense)	(None, 50)	2550
dense_27 (Dense)	(None, 50)	2550
dense_28 (Dense)	(None, 1)	51

Total params: 94,401
Trainable params: 94,401
Non-trainable params: 0

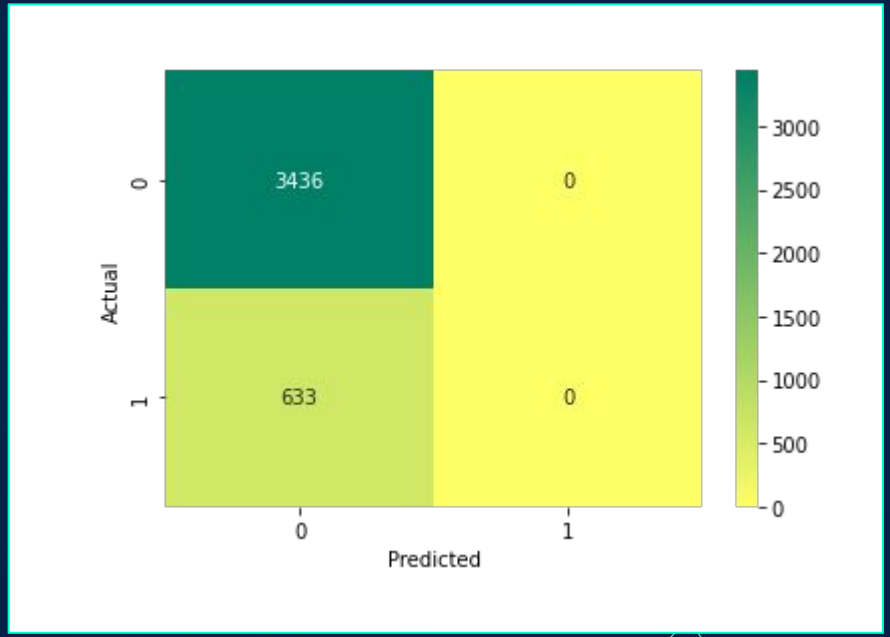




Multilayer Perceptron-2

	precision	recall	f1-score	support
0	0.84	1.00	0.92	3436
1	0.00	0.00	0.00	633
accuracy			0.84	4069
macro avg	0.42	0.50	0.46	4069
weighted avg	0.71	0.84	0.77	4069

We can see the MLP failed due to the imbalances.



Model Comparison

	Logistic Regression	Naive Bayes	LightGBM	Multilayer Perceptron
Training Accuracy	0.92	0.80	0.89	0.50
Testing Accuracy	0.70	0.75	0.87	0.84

best performing model

We define accuracy rate as our evaluation metrics since accuracy rate is most important in our business case.

05

Conclusion & Evaluation

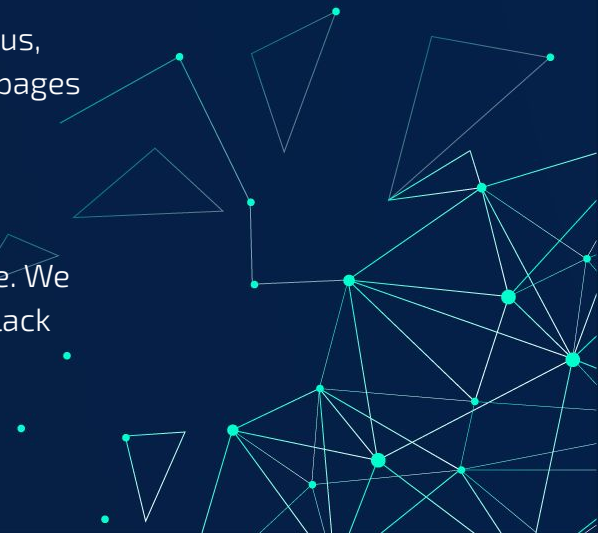
Conclusion

Our team managed to train and compared 4 different models and selected the lightGBM model as our final choice. LightGBM, a gradient boosting tree model, predicts whether a customer will place order online with a 0.89 accuracy rate on training set and 0.87 accuracy on testing set.

Insights

In addition, the model indicates that certain types of page customers viewed (Administrative and Product Related) tend to best predict if they will purchase. Thus, we recommend the company put efforts to design the contents of these types of pages and put more marketing campaign to attract online customers

From time perspective, we realize that customers visiting the website during November and December has higher importance in predicting if they will purchase. We believe such finding reflects real life because these two months are periods for Black Friday and Christmas respectively.



Evaluation

Economic Assumptions

1. Assume after the model identifying if a customer will purchase, the company will apply several marketing strategies (recommend products, promotion code, etc).
2. Such campaign will increase 20% order value for customers that will purchase initially and convert 10% customers who initially will not purchase.
3. The average order value is $128 according to statistics.
4. Also, the annual growth for online visitor is 10%.

Economic Impact

Place Order	Customers (Current)	Customers (Next Year)	Additional Impact
Yes	1908	2098	$ 53,709
No	10422	11464	$ 146,739
Total	12330	13562	**$ 200,448**

06

Future Work

Future Improvement

Data Collection

Suggest that the company provides data for the missing months and collect more data on successful purchase to better train our models

Feature Mining

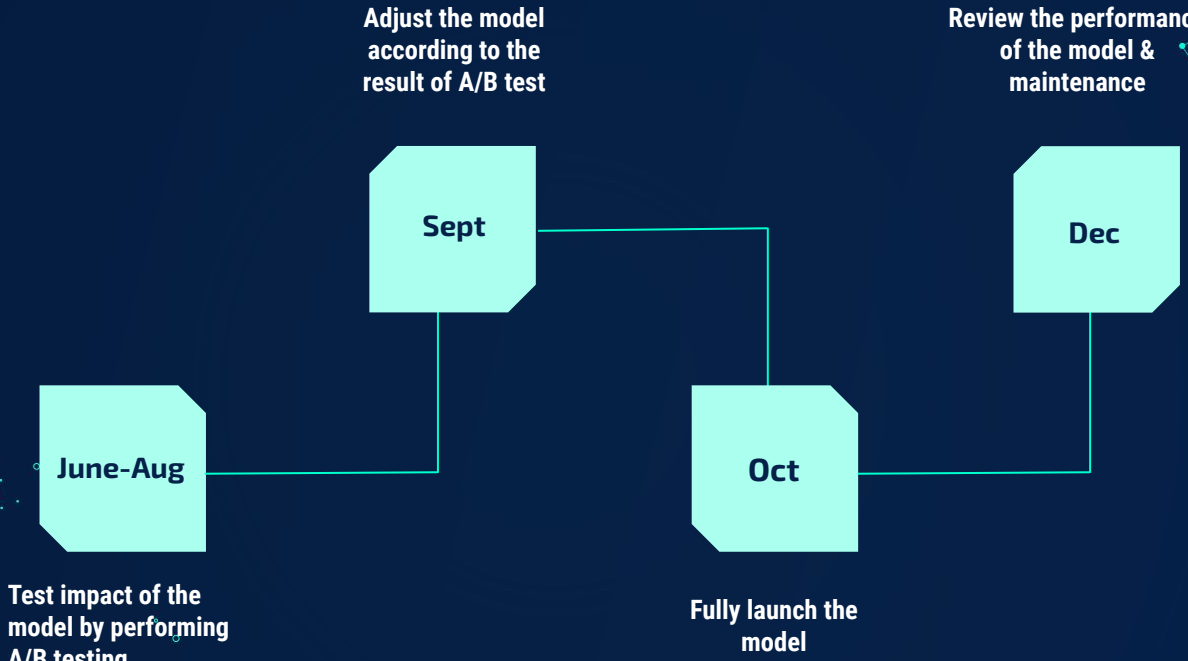
Deep dive into feature engineering to include more features into the model and improve model performance

Modeling

Iterate through different combinations for MLP (activation function, number of layer, number of neurons) to find the best model for prediction

Also, applying grid search and cross validation techniques to fine tune our model can be another improvement

Recommendation



Reference

10 Ecommerce Average Order Value Statistics (Updated 2020) (growcode.com)

THANKS

2nd ANNOUNCEMENT

MARS

Despite being red, Mars is a cold place, not hot

NEPTUNE

Neptune is the farthest planet from the Sun, the fourth-largest in our Solar System

JUPITER

It's a gas giant and the biggest planet in our Solar System

SATURN

Saturn is a gas giant, composed of hydrogen and helium

IN DEPTH

MERCURY

Mercury is the closest planet to the Sun and the smallest one in the Solar System—it's only a bit larger than our Moon. The planet's name has nothing to do with the liquid metal, since it was named after the Roman messenger god, Mercury

OUR NUMBERS

500 new clients last semester

100 ongoing projects

40 new employees last semester

NEWS

NEPTUNE

Neptune is the farthest planet from the Sun and the fourth-largest in our Solar System

MARS

Despite being red, Mars is a cold place, not hot. It's full of iron oxide dust, which gives the planet its reddish cast

MERCURY

Mercury is the closest planet to the Sun and also the smallest one in our Solar System

JUPITER

Jupiter is the biggest planet in our Solar System and also the fourth-brightest object in the sky

VENUS

Venus has a beautiful name and is the second planet from the Sun. It's terribly hot, even hotter than Mercury

SATURN

Saturn is the ringed planet. It's a gas giant, composed mostly of hydrogen and helium

04

WELCOME!

You can enter here the subtitle if you need it

WELCOME!

MARY RUIZ

Neptune is the farthest planet from the Sun and the fourth-largest in our Solar System

JOHN SMITH

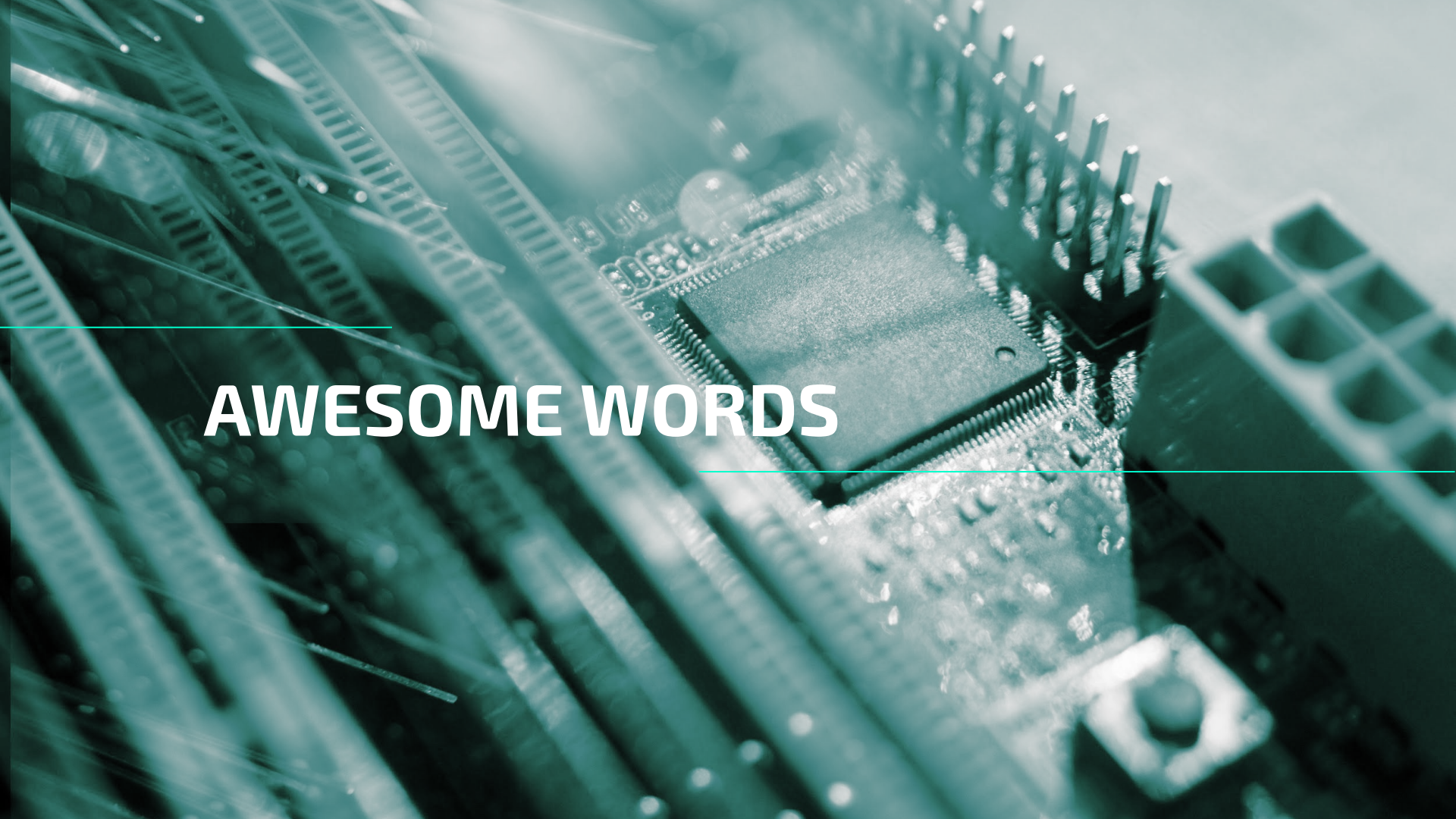
Despite being red, Mars is a cold place, not hot. It's full of iron oxide dust, which gives the planet its reddish cast

JONATHAN DOE

Mercury is the closest planet to the Sun and the smallest one in our Solar System

MINA HAWKINS

Venus has a beautiful name and is the second planet from the Sun. It's terribly hot, even hotter than Mercury


AWESOME WORDS

UPCOMING EVENTS

SEP 08

Mercury is the closest planet to the Sun

OCT 16

Despite being red, Mars is a cold place, not hot

NOV 29

Saturn is composed of hydrogen and helium

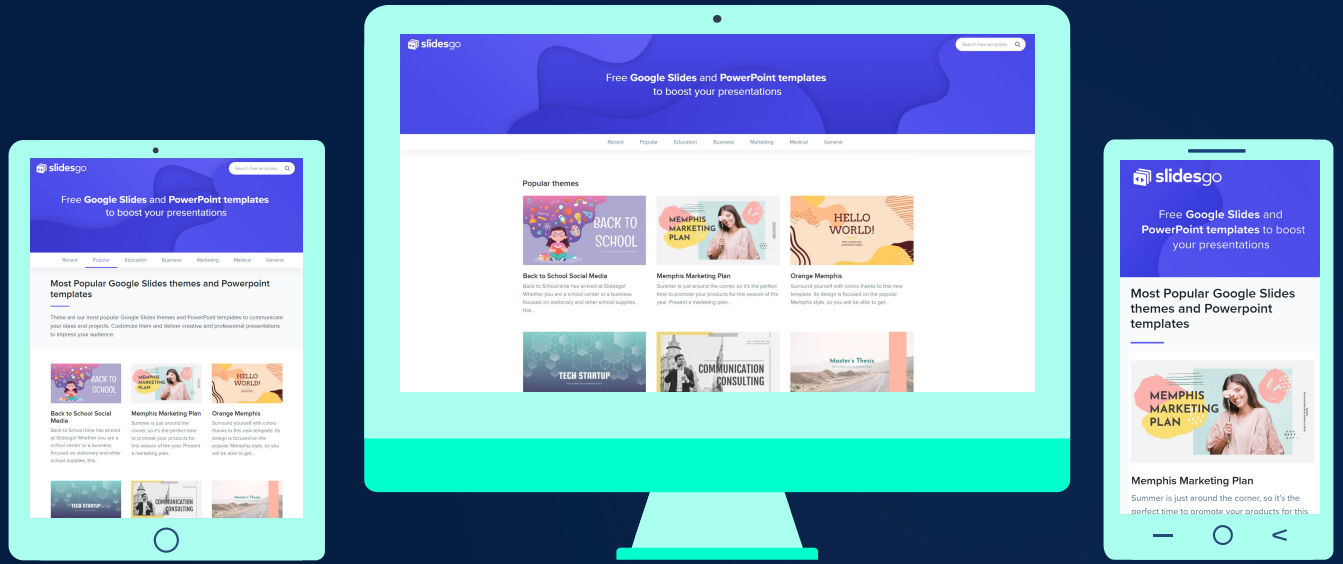
DEC 10

Jupiter is the biggest planet in our Solar System

08

Recommendations

SNEAK PEEK



Neptune is the farthest planet from the Sun and the fourth-largest in our Solar System

SNEAK PEEK

	Administrative	Administrative_Duration	Informational	Informational_Duration	ProductRelated	ProductRelated_Duration	BounceRates	ExitRates
0	0	0.0	0	0.0	1	0.000000	0.20	0.20
1	0	0.0	0	0.0	2	64.000000	0.00	0.10
2	0	0.0	0	0.0	1	0.000000	0.20	0.20
3	0	0.0	0	0.0	2	2.666667	0.05	0.14
4	0	0.0	0	0.0	10	627.500000	0.02	0.05

PageValues	SpecialDay	Month	OperatingSystems	Browser	Region	TrafficType	VisitorType	Weekend	Revenue
0.0	0.0	Feb	1	1	1	1	Returning_Visitor	False	False
0.0	0.0	Feb	2	2	1	2	Returning_Visitor	False	False
0.0	0.0	Feb	4	1	9	3	Returning_Visitor	False	False
0.0	0.0	Feb	3	2	2	4	Returning_Visitor	False	False
0.0	0.0	Feb	3	3	1	4	Returning_Visitor	True	False

Instructions for use (free users)

In order to use this template, you must credit **Slidesgo** by keeping the Thanks slide.

You are allowed to:

- Modify this template.
- Use it for both personal and commercial purposes.

You are not allowed to:

- Sublicense, sell or rent any of Slidesgo Content (or a modified version of Slidesgo Content).
- Distribute this Slidesgo Template (or a modified version of this Slidesgo Template) or include it in a database or in any other product or service that offers downloadable images, icons or presentations that may be subject to distribution or resale.
- Use any of the elements that are part of this Slidesgo Template in an isolated and separated way from this Template.
- Delete the “Thanks” or “Credits” slide.
- Register any of the elements that are part of this template as a trademark or logo, or register it as a work in an intellectual property registry or similar.

For more information about editing slides, please read our FAQs or visit Slidesgo School:

https://slidesgo.com/faqs and https://slidesgo.com/slidesgo-school

Instructions for use (premium users)

In order to use this template, you must be a Premium user on **Slidesgo**.

You are allowed to:

- Modify this template.
- Use it for both personal and commercial purposes.
- Hide or delete the “Thanks” slide and the mention to Slidesgo in the credits.
- Share this template in an editable format with people who are not part of your team.

You are not allowed to:

- Sublicense, sell or rent this Slidesgo Template (or a modified version of this Slidesgo Template).
- Distribute this Slidesgo Template (or a modified version of this Slidesgo Template) or include it in a database or in any other product or service that offers downloadable images, icons or presentations that may be subject to distribution or resale.
- Use any of the elements that are part of this Slidesgo Template in an isolated and separated way from this Template.
- Register any of the elements that are part of this template as a trademark or logo, or register it as a work in an intellectual property registry or similar.

For more information about editing slides, please read our FAQs or visit Slidesgo School:

https://slidesgo.com/faqs and https://slidesgo.com/slidesgo-school

Fonts & colors used

This presentation has been made using the following fonts:

Exo 2
(https://fonts.google.com/specimen/Exo+2)

Roboto Condensed
(https://fonts.google.com/specimen/Roboto+Condensed)

#092a5c

#00ffcd

#ffffff

Storyset

Create your Story with our illustrated concepts. Choose the style you like the most, edit its colors, pick the background and layers you want to show and bring them to life with the animator panel! It will boost your presentation. Check out How it Works.

Pana

Amico

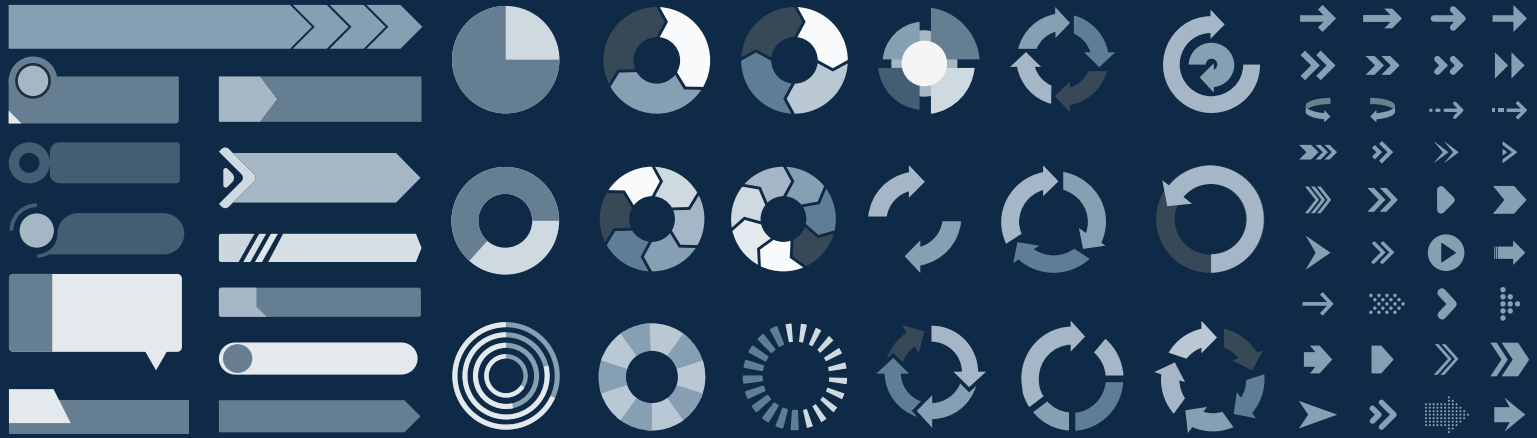
Bro

Rafiki

Cuate

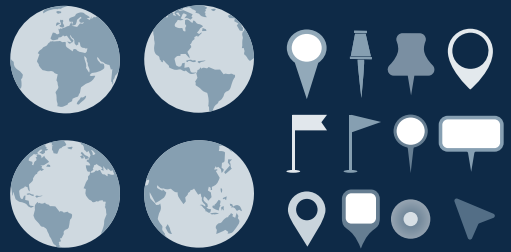
Use our editable graphic resources...

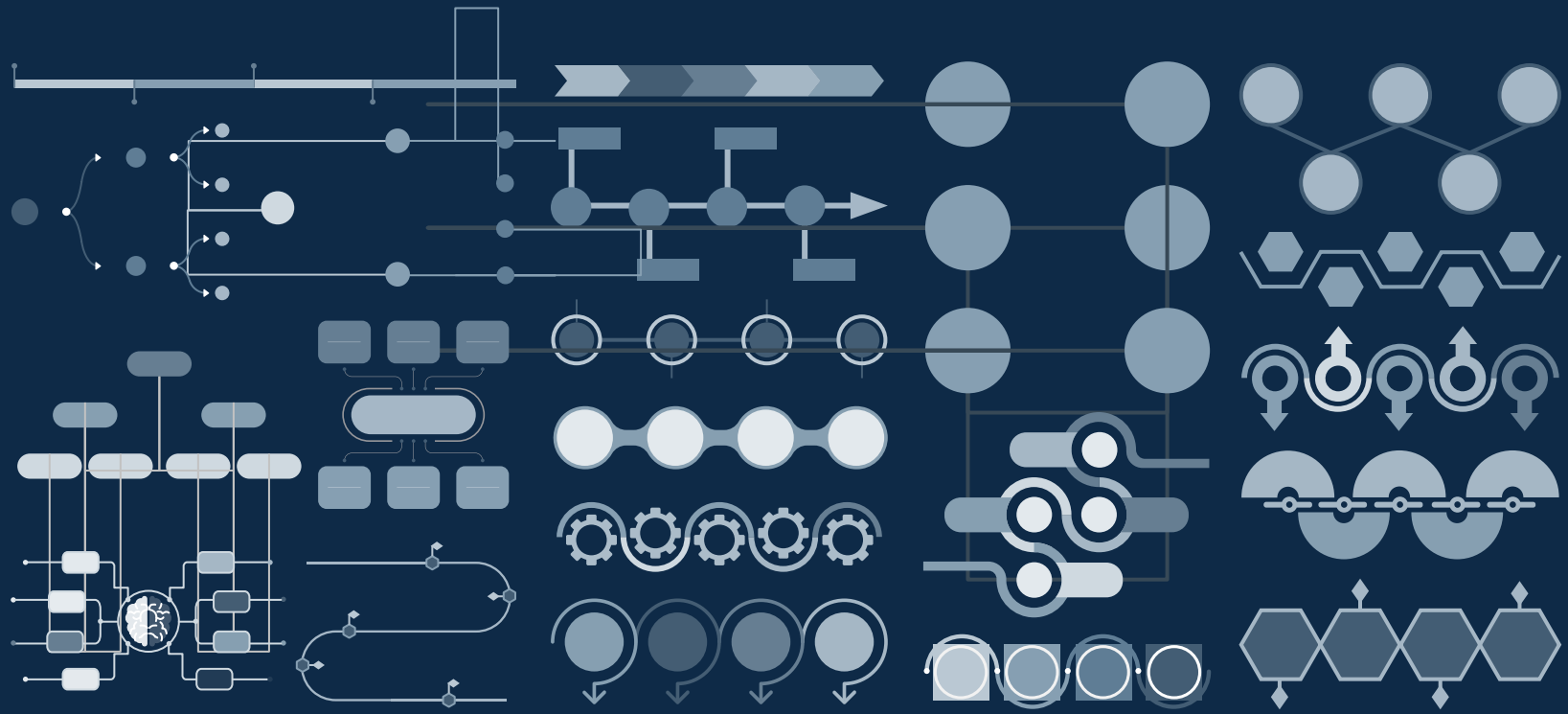
You can easily resize these resources without losing quality. To change the color, just ungroup the resource and click on the object you want to change. Then, click on the paint bucket and select the color you want. Group the resource again when you're done. You can also look for more infographics on Slidesgo.

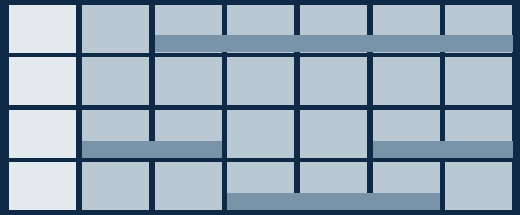


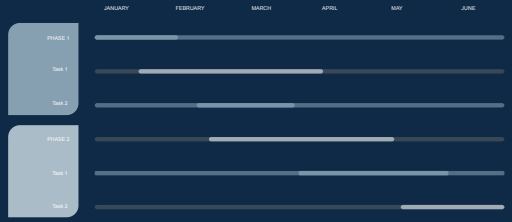


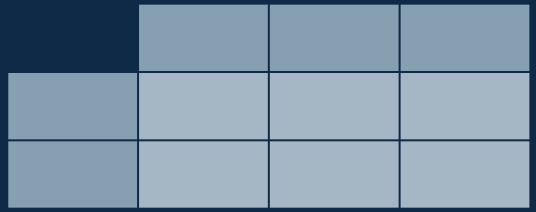


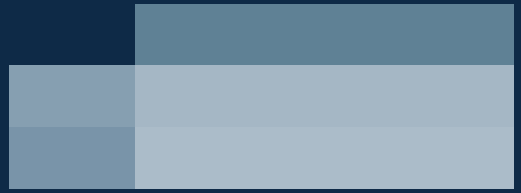



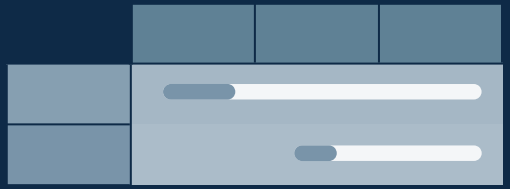


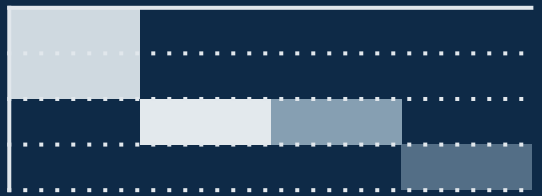


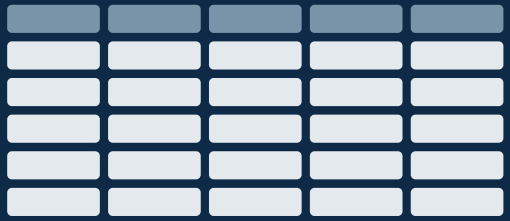

JANUARY
FEBRUARY
MARCH
APRIL
MAY
JUNE
PHASE 1
Task 1
Task 2
PHASE 2
Task 1
Task 2

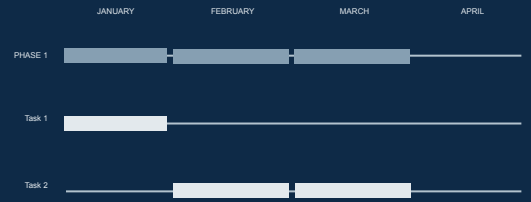


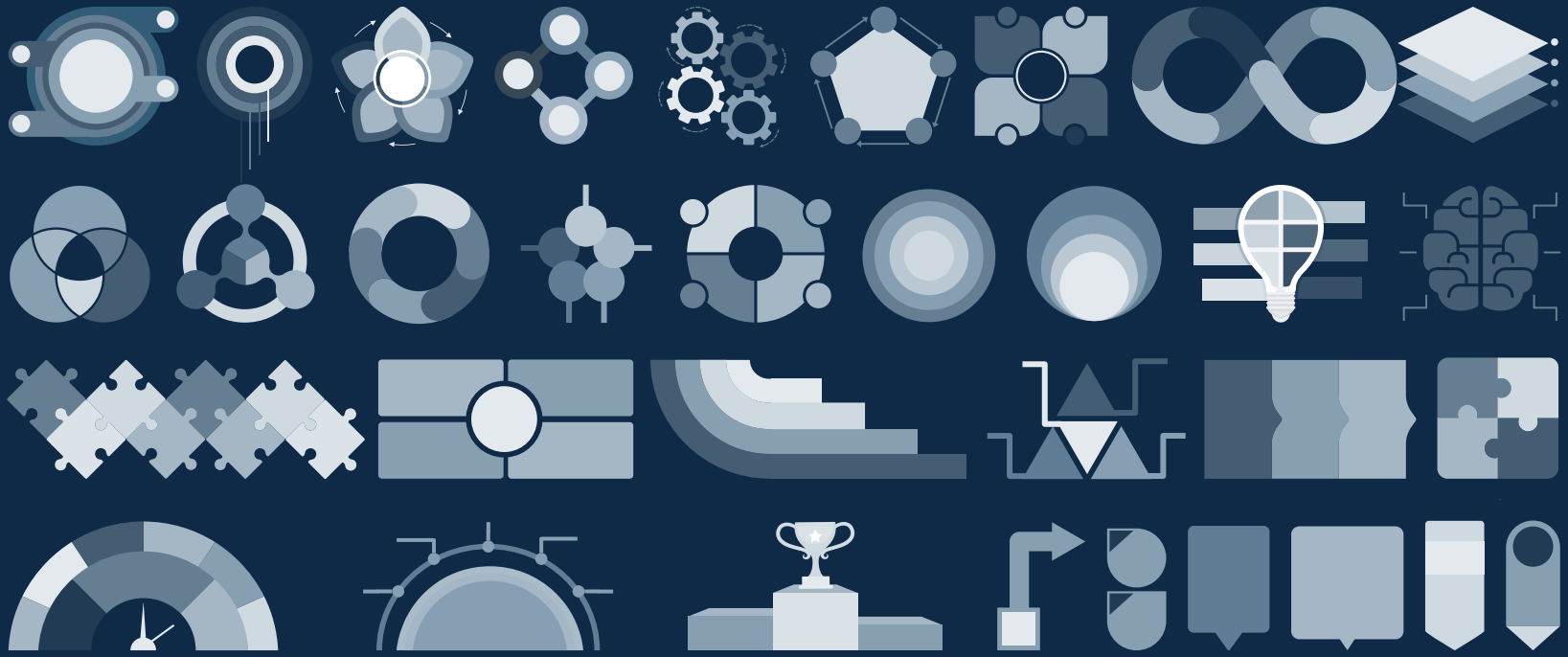


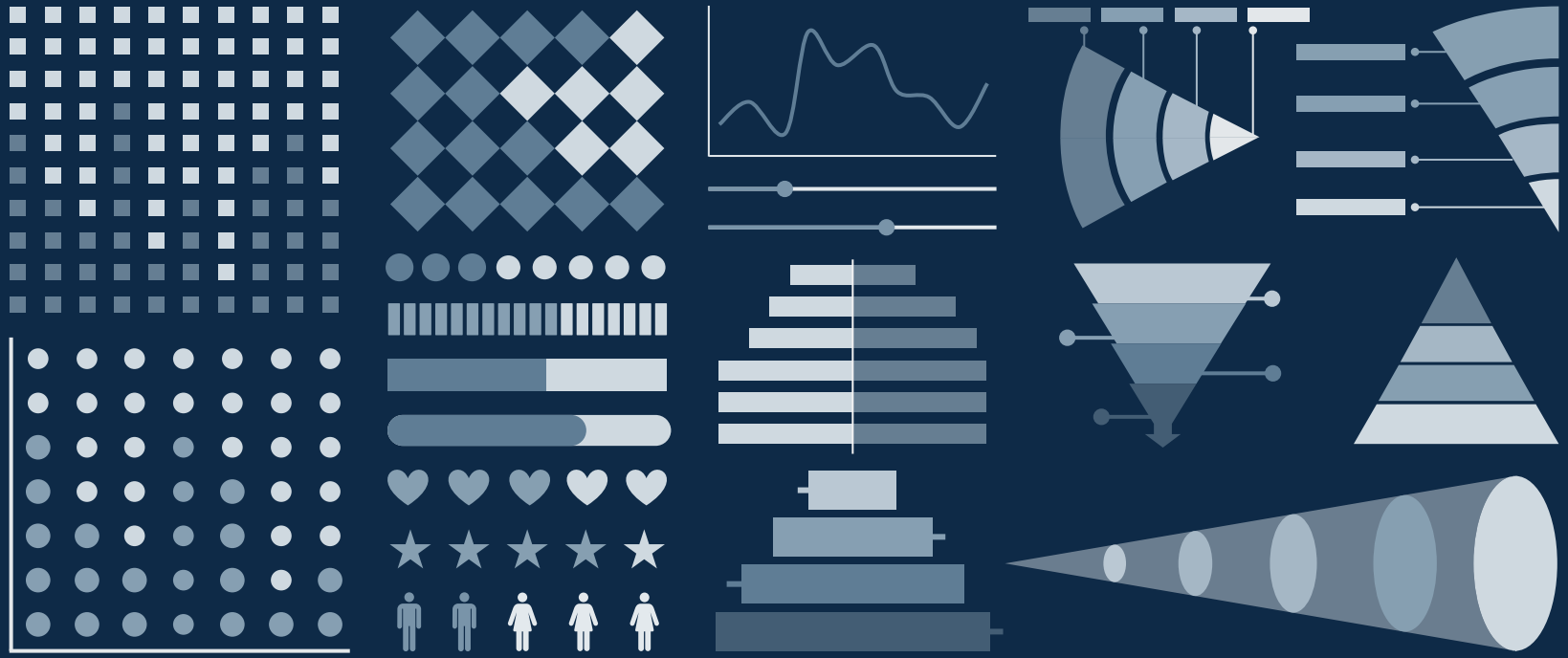







JANUARY
FEBRUARY
MARCH
APRIL
PHASE 1
Task 1
Task 2





...and our sets of editable icons

You can resize these icons without losing quality.
You can change the stroke and fill color; just select the icon and click on the paint bucket/pen.
In Google Slides, you can also use Flaticon's extension, allowing you to customize and add even more icons.

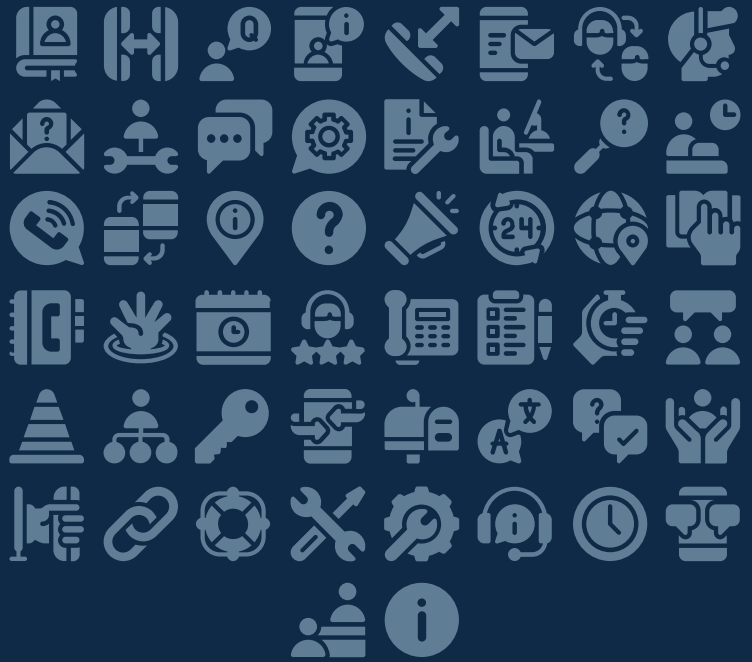
Educational Icons

Medical Icons

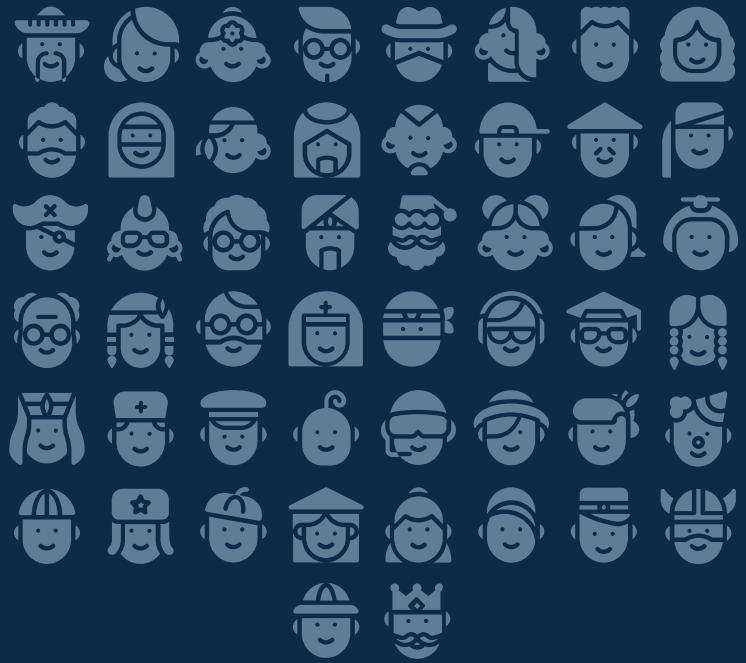
Business Icons

Teamwork Icons

Help & Support Icons



Avatar Icons



Creative Process Icons

Performing Arts Icons

Nature Icons

SEO & Marketing Icons


slidesgo