CUSTOMER REVENUE PREDICTION

Finding the needle in the haystack

TABLE OF CONTENTS

01 **About the Project**
Understand the task

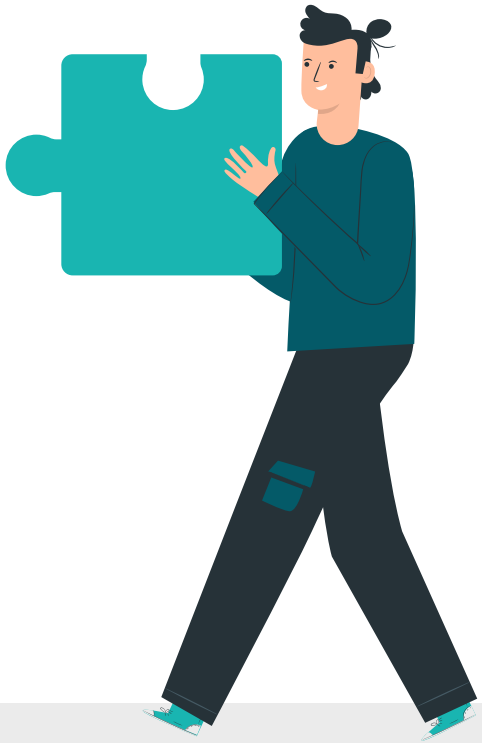
02 **Insights**
Interesting and funny things

03 **Model**
How to solve the problem

04 **Business Use**
How the model can be used in a business case

05 **Key Takeaways**
Some well-meant advice

06 **Future Work**
There is a lot more to look at



ABOUT THE PROJECT

OVERVIEW

01

kaggle

Kaggle is a platform for competitions to solve data science challenges

02

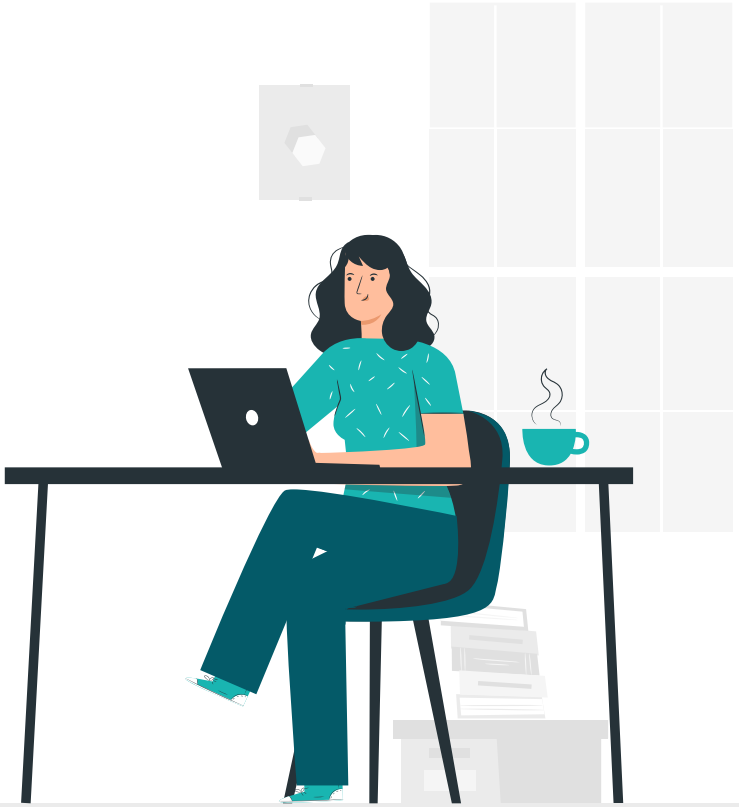
Google Merchandise Store

An online shop where Google merchandise is sold

03

Google Analytics

Google Analytics is a web analytics service that tracks and reports website traffic



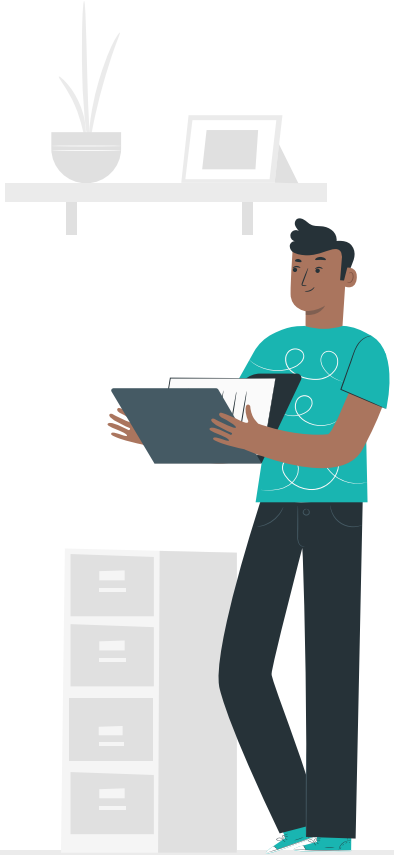
THE TASK

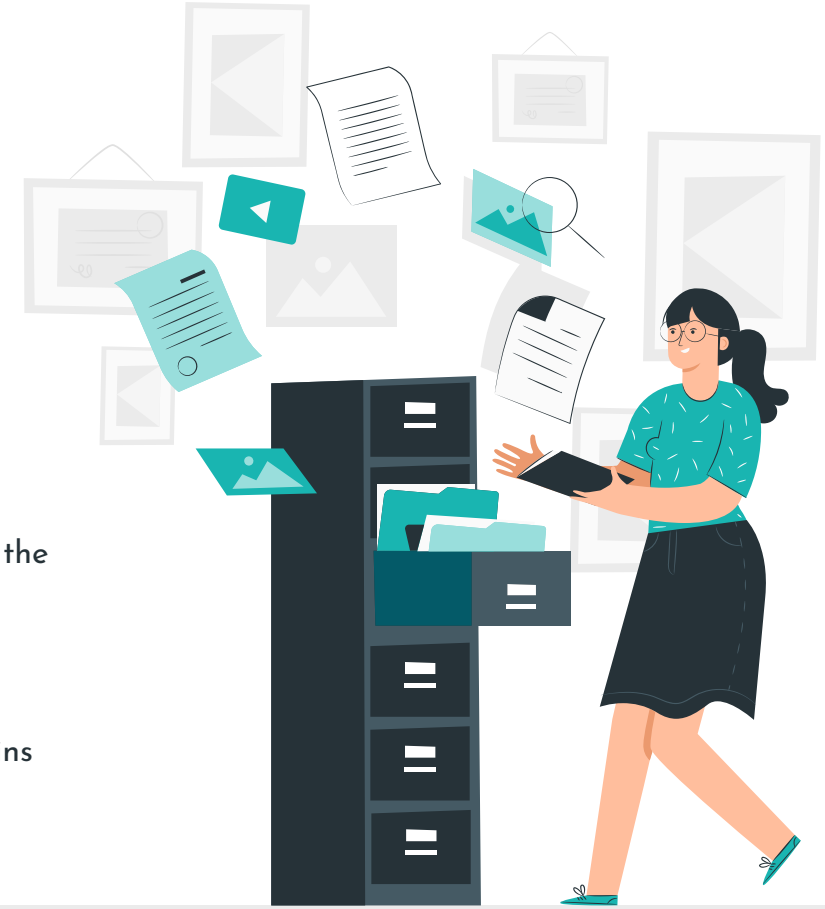
We are challenged to analyze a **Google Merchandise Store** customer dataset to predict revenue per customer.

Firstly we are given a data training set containing user transactions from **August 1st 2016 to April 30th 2018** to train our model.

Secondly we are given a test data set from **May 1st 2018 to October 15th 2018** for the prediction.

We are predicting how much money a user from the test data set will spend in the unseen period of **December 1st 2018 to January 31st 2019.**



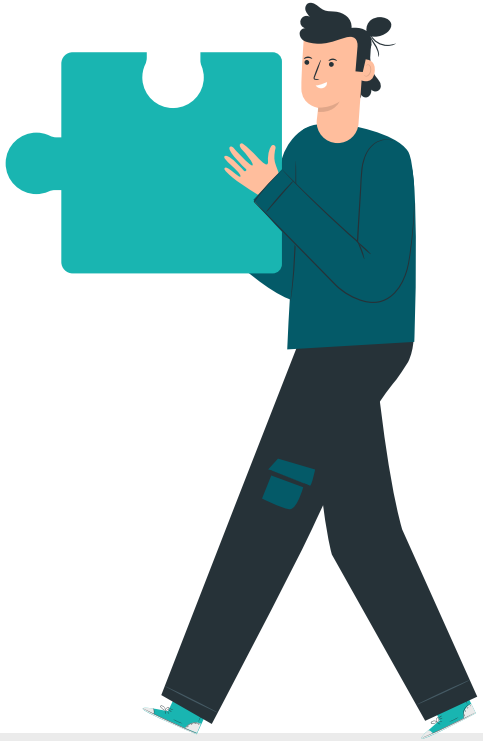


CHALLENGES

31 GB data set: It takes a good while to import the data before one can even start editing.

149 features: Understanding the importance of each feature and the circumstances in which missing values occur.

Needle in a haystack: The test data set contains almost 300,000 IDs, based on the train data, probably only about 40 IDs are the returning buyers to be found.



INSIGHTS

SOME DAYS WITH EXTREMELY HIGH REVENUES

Checking the data revealed: There was one single person who spent $30,170 on the 26th July 2018. The other peaks are also caused by individuals.
Assumption: Companies buy for their employees and/or as promotional gifts.

First Day: 08/01/16
Last Day: 10/15/18

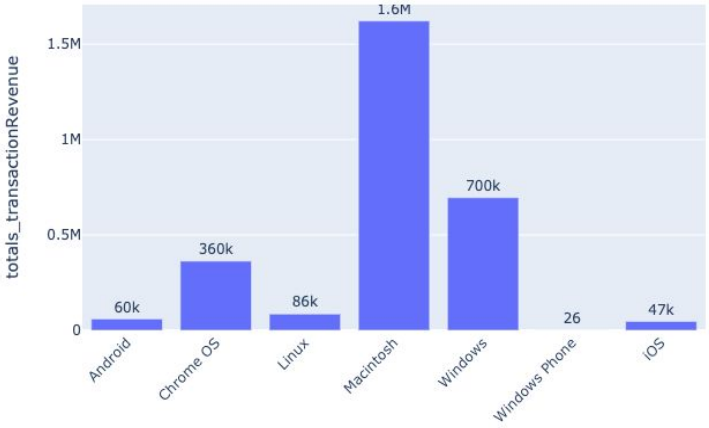
Avg. revenue per day: $3,567

Revenue: 07/26/18
$34,138

WHERE IS THE MONEY?

$2.9 millions in total

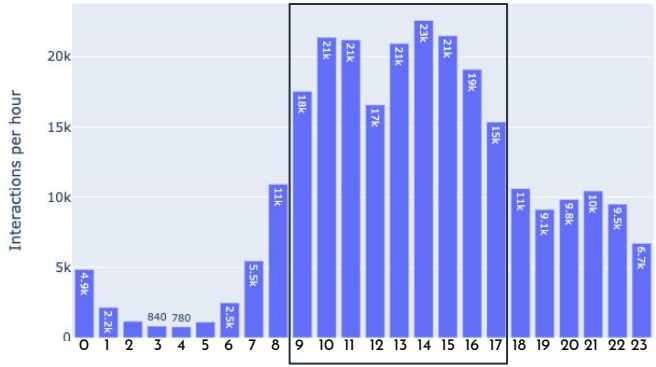
Mac users account for 56% of total sales



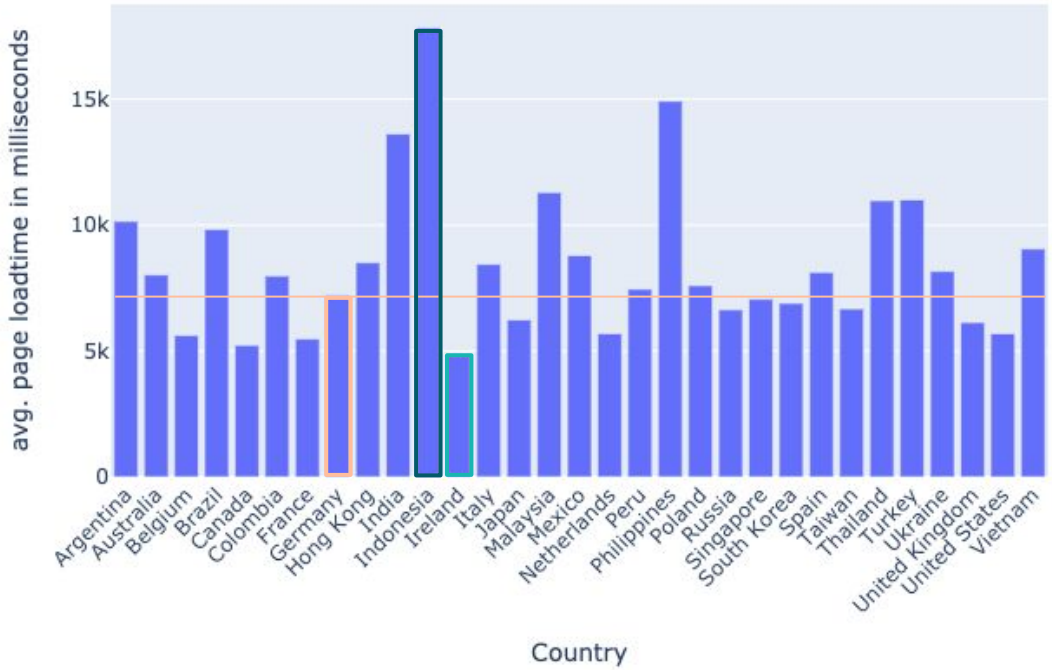
SHOPPING 9 TO 5

Lunch at 12

It seems that office hours and website traffic are highly correlated



WHERE TO FIND THE FASTEST INTERNET?



Germany

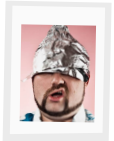
Not as bad as I thought

Ireland

Fast internet seems to be no problem here

Indonesia

They still have a long way to go



?

NEVER TRUST A PENGUIN – THERE COULD BE A CONNECTION!

VISITORS FROM ALL OVER THE WORLD

43%

Of all visitors come from the US

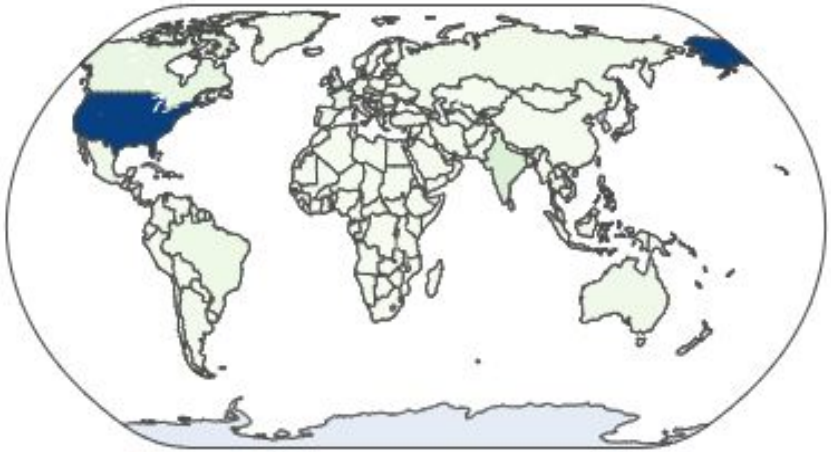
0

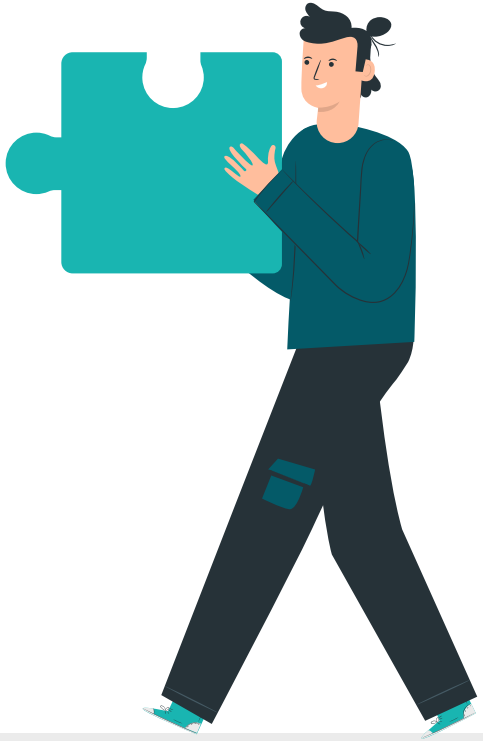
From North Korea

0

From Antarctica

There were shop visitors from all over the world, but Antarctica and North Korea are missing!





MODEL

WHAT IS A REALISTIC RESULT (RMSE)?

Kaggle winner

The best result from over a thousand submissions

Baseline model

The result to beat: Predicting all revenues with "zero", which would have been place 304 out of 1,049 in the competition

My model

Somewhere in between (fun fact: Most people even failed to beat the baseline model) So my goal is to do better!

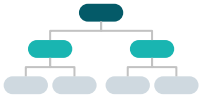
MODEL STEP BY STEP

A lot of work to get the data set ready

PREPARATION

The IDs predicted as returning buyers (5!) receive a possible revenue value via a regression model

REGRESSION



BASELINE

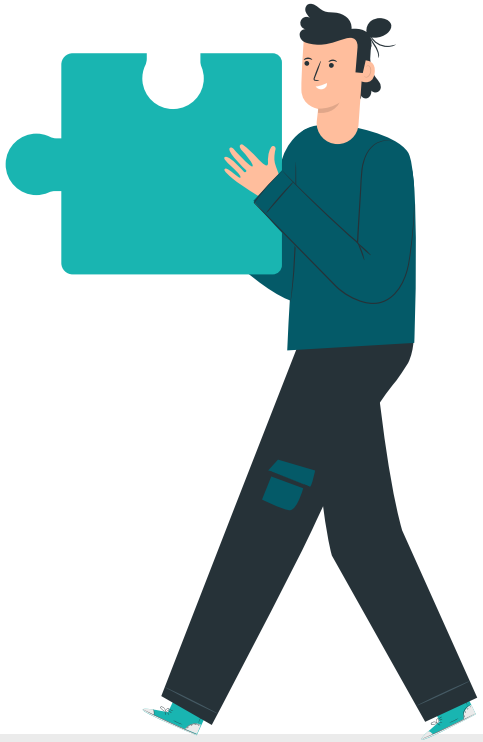
RMSE = 0.88843
Place 304

CLASSIFICATION

I used LightGBM to predict whether or not a buyer will buy again in the specified time window

FINAL RESULT

RMSE = 0.8828 ✓
Place 140



BUSINESS USE

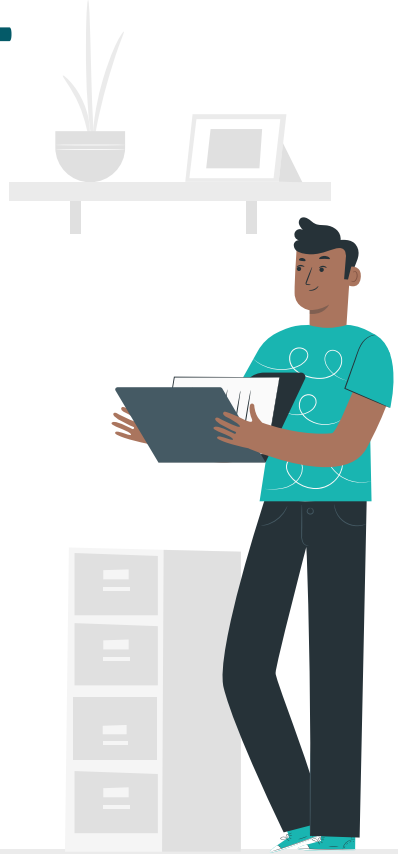
INCREASE SALES

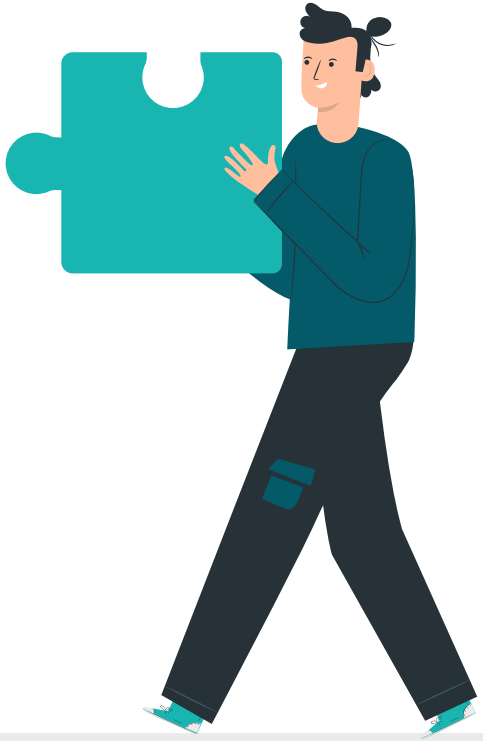
If you use advertising to influence sales, you need a little more than just five people who will order again. With a small adjustment the model can be used for marketing purposes. The goal would most likely be, to get as many people as possible to buy in your store. The most efficient way is to focus your marketing budget on people with the highest probability to buy.

HOW TO ADJUST THE MODEL

We use the existing model and apply the probability for the classification as buyer. Depending on our marketing budget, we could select the 10,000 IDs (as an example) with the highest purchase probability and use them for advertising.

	fullVisitorId	probability
191200	6451020629616527625	78.643809
239323	8073822829065741671	74.086966
243467	8216311071672550835	61.799112





KEY TAKEAWAYS

KEY TAKEAWAYS

01

Customers

Know as much as possible about your customers to develop new sales strategies

02

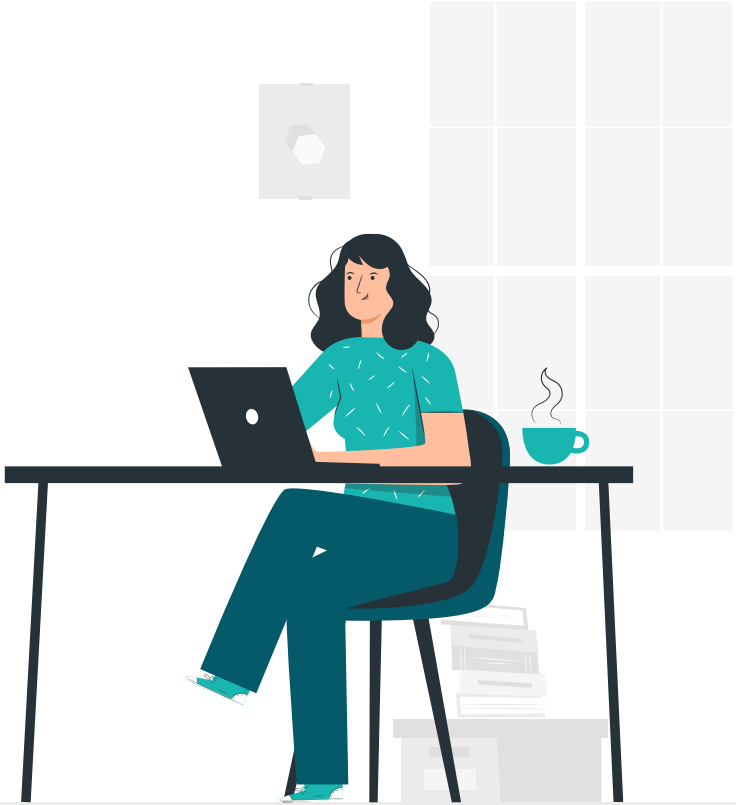
Data

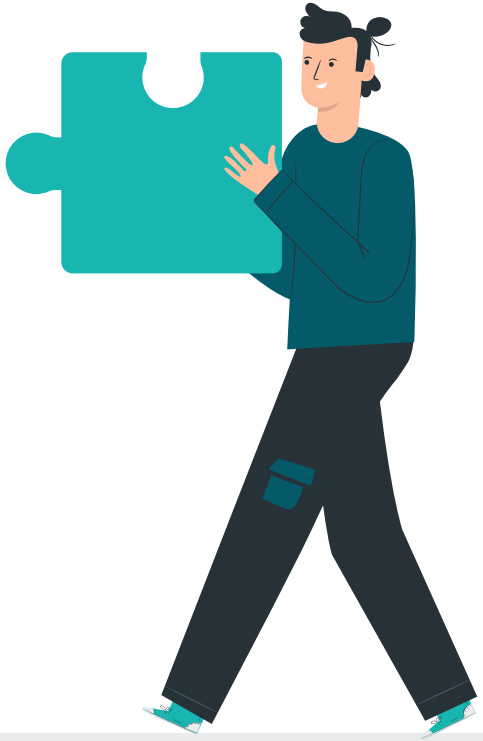
Get more data to achieve a higher forecast quality

03

Penguins

They could be evil, better be safe than sorry





FUTURE WORK

FUTURE WORK

EDA

Get more insights out of the data set

Regression

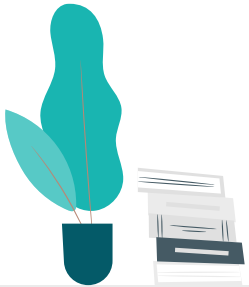
Use a more advanced model for regression

Local Time

Could the model be improved by using the local time

Products

Try to find a way to predict the product




I NEED A BEER
NOW