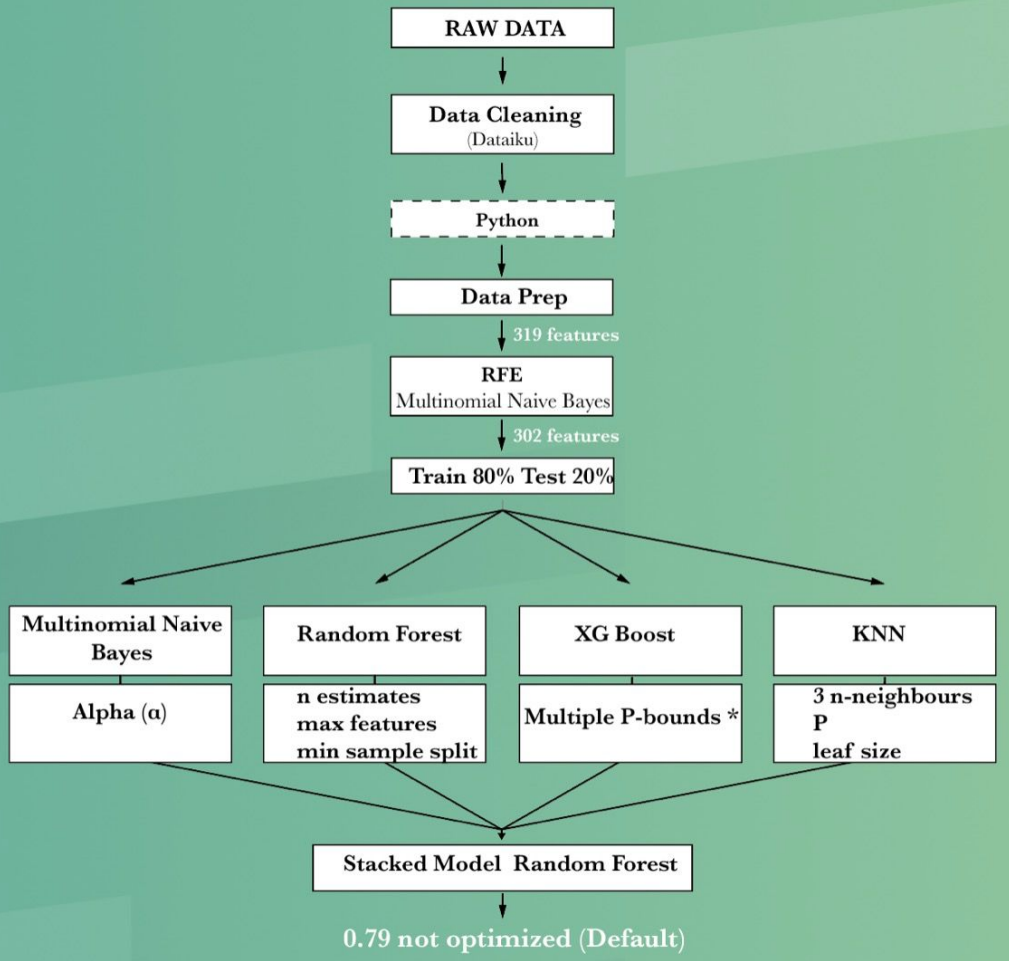
Pump it Up: Data Mining the Water Table

Machine Learning II

MBD OCT 2018 - 01-7 - Group G

1.
Project Structure


RAW DATA
Data Cleaning
(Dataiku)
Python
Data Prep
319 features
RFE
Multinomial Naive Bayes
302 features
Train 80% Test 20%
Multinomial Naive Bayes
Random Forest
XG Boost
KNN
Alpha (α)
n estimates
max features
min sample split
Multiple P-bounds *
3 n-neighbours
P
leaf size
Stacked Model Random Forest
0.79 not optimized (Default)

2.
Data Preparation and Feature Engineering

Data Preparation and Feature Engineering

Feature Engineering: Concatenation of Geographic Features and Creation of New Features

- 'Region' and 'District' were concatenated to uniquely identify each district
- The wells with government funder or installer were identified in new features

Feature Engineering: Creation of New Features through and External Dataset

1

- Join with coordinates of centre of gravity of the water basins
- Compute distance in km. between the well and the water basin

Data Preparation

3

- Imputing null values
- Removing duplicate columns
- Removing columns with an excessive number of levels

3.
Data Exploration Key Insights

Exploratory Data Analysis

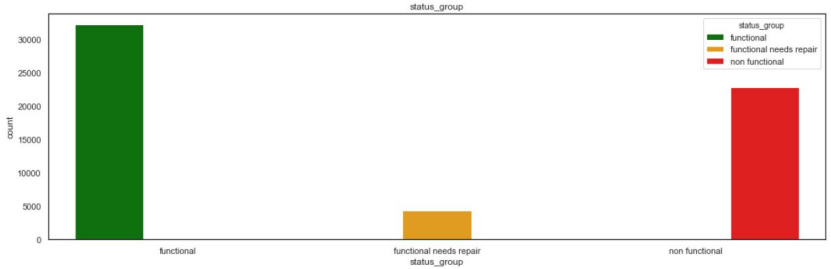


Figure 1: ‘Status_Group’ Variable Barchart

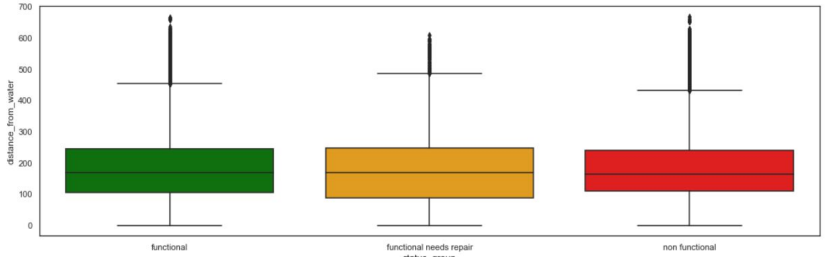


Figure 2: ‘Distance_from_water’ Boxplot

Target Variable Unbalanced Dataset

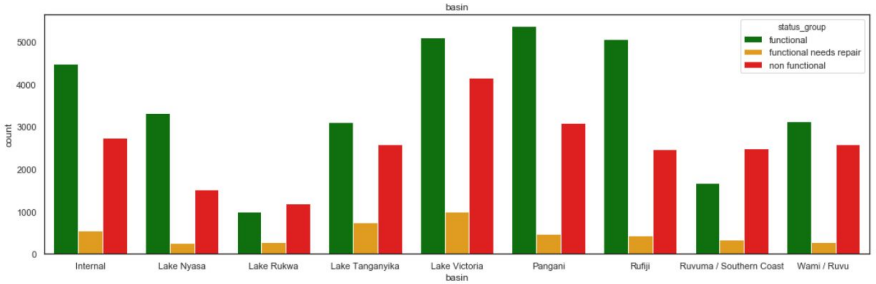


Figure 3: ‘Basin’ Variable Barchart

Exploratory Data Analysis

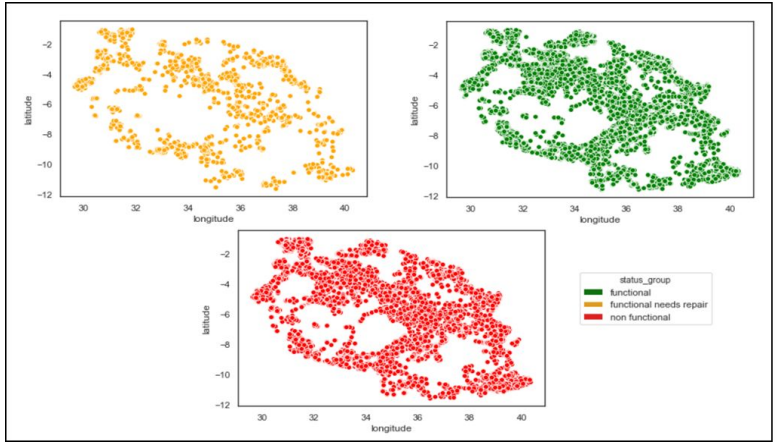


Figure 4: Coordinates Scatter Plot by Target Outcome

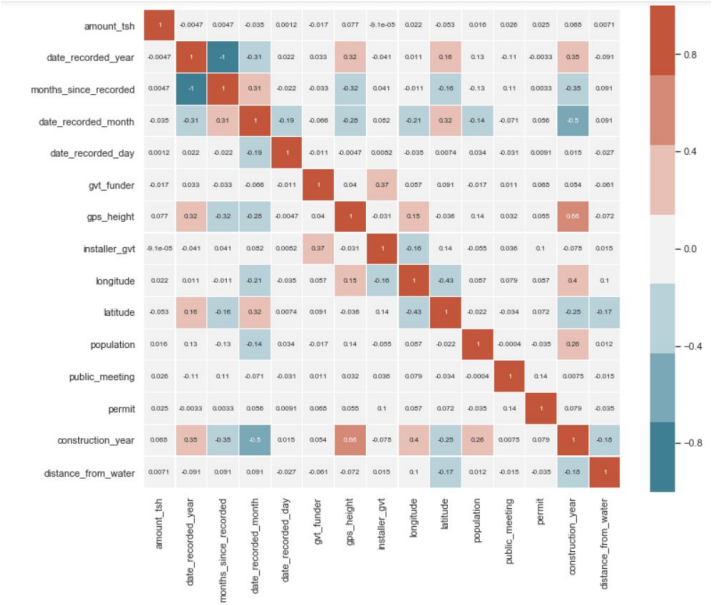


Figure 5: Confusion Matrix

4.
Baseline Model

Baseline Model

- Logistic Regression Model
- ‘L1’ Penalty Applied
- Balanced Data Set

Accuracy
0.714

BASELINE

329
Features

5.
Feature Selection

Feature Selection

- Recursive Feature Elimination (RFE) with Naive-Bayes (Multinomial) estimator
- Optimized for number of features yielding highest accuracy

Principal Component Analysis (PCA) yields only 1 PC which explains >99% of the features but is not usable for actual predictions (due to low accuracy scores)

Features selected

302 Features

6.
Other Models

Multinomial Naive Bayes

Why Naive Bayes?

- Fast and simple
- Dimensionality not an issue

Parameter	How it was Determined
Alpha: 25.7%	Bayes Optimization- parameters (10,30)
RESULT	**68.57%**

KNN

Why KNN?

- Easy to program
- Classification Accuracy can be very good
- Optimal for models with < 20 features, however built for distance based data

Parameter	How it was Determined
K/Number of Neighbors:15	Bayes Optimization- parameters (5, 20)
Distance Measure: Minkowski	Default
p:1	Bayes Optimization- parameters (1,2)
Leaf Size: 29	Bayes Optimization- parameters (20,40)
RESULT	**77.37%**

XGBoost

Why XGBoost?

- Classification Accuracy can be very good
- Ensemble learning method
- Reduces bias and variance

Parameter	How it was Determined
Learning Rate	Weighting factor for correction on new trees, to slow down learning rate: (0.2, 0.5)
Gamma	Not using high depth in the case: (0, 1)
Maximum Depth	Maximum Depth of the trees: (5, 20)
Minimum Child Weight	Controls the pruning of the derivative: (0.8, 2)
Maximum Delta Step	Constrains the maximum weight given to any particular tree: (0, 10)
Subsample	Fractions of observations to be sampled from each tree: (0.5, 1)
Column Sample by Tree	Fractions of columns that can be assessed with a particular tree: (0.5, 1)
Regular Lambda	Regularization (L2) constraints on weights: (0.5, 1.5)
Regular Alpha	Regularization (L1) constraints on weights:(0, 1)
RESULT	**78.68%**

Random Forest Classifier

Why Random Forest?

- Less chance for over-fitting
- Accurate and robust
- Good for large number of features

Parameter	How it was Determined
N_estimators: 115	Bayesian Optimization (10,250)
Min_samples_split: 16	Bayesian Optimization (2,25)
Max_features: .272	Bayesian Optimization (0.1,0.999)
RESULT	**79.99%**

7.
Final Model & Conclusions

Stacked Model - Random Forest Classifier

- Multi-class (label-encoded) features corresponding with target (same scale)
- Suited for same prediction/feature values corresponding to different target value
- Accurate and robust

We also added Polynomial Features (basis stack dataset) in pipeline before RFC.

Parameter	How it was Determined
N_estimators: 180	Bayesian Optimization (10,250)
Min_samples_split: 25	Bayesian Optimization (2,25)
Max_features: 0.1412	Bayesian Optimization (0.1,0.999)

Optimized through Bayesian-Optimization technique

Stacking Flow

Initial Models	Target Variable Prediction with Initial Models	Stacked Model (RFC)
Use the entire training set to fit the 4 primary models - KNN, MNB, RFC, XGB, in order eventually run predictions on the test set.	Predict the target using 5-fold cross prediction method, to be used for training the stacked (final) model.	Based on the information from the 5-fold cross prediction and 2 features (longitude and latitude), fit the stacked model (optimized using Bayesian Optimization) for future prediction using the the hold-out/test set

Stacking Flow

Import and Preparation of Hold-out Set	Predict and Creation DF with Initial Models	Predict with Stacked Model (RFC)
Import and transform the hold-out set - including feature engineering and feature selection (RFE). This ensures that the dimensions of the fitted models and hold-out set are same.	Based on the fitted models (training set), we predict the target for the hold-out set and create the stacked model.	Based on the fitted stacked model (training set), create final prediction for the target in the hold-out set to complete the final model.

Final Metrics

78.38%

Accuracy on hold-out set for Stacked Model

80.46%

Accuracy on hold-out set for Random Forest Model

Optimization still required:

- Add constructed features to the stacked model (i.e. predictions from primary models and target) in order to differentiate predictions from primary models for the added features- this adds a new and necessary level of complexity to the model, but also allows for some flexibility.
- Run model without the underperforming MNB (replace with LDA/ Log Reg), which had accuracy was 10 points below the three other primary models

Baseline:	**0.76**	**57**
Day:	**0.76**	**88**
Month-Day:	**0.75**	**423**
Peaks:	**0.87**	**58**
Temp(x4):	**0.87**	**59**
Polynomials:	**0.87**	**64**
Hours Bins:	**0.83**	**40**
RFE:	**0.87**	**55**
RFE Lite:	**0.82**	**36**
Manual:	**0.85**	**46**
Manual+Rain:	**0.86**	**50**
Manual+FW:	**0.86**	**53**

SLIDES BELOW ARE ONLY TEMPLATES

TO COPY AND USE IN THE MAIN PRESENTATION ONLY

KNN

Reason:

Although model is optimal for usual

Technologies

Is the colour of the clear sky and the deep sea. It is located between violet and green on the optical spectrum.

Future Usage

Is the color of blood, and because of this it has historically been associated with sacrifice, danger and courage.

Context

More info on how to use this template at www.slidescarnival.com/help-use-presentation-template

EDIT IN GOOGLE SLIDES

Click on the button under the presentation preview that says **"Use as Google Slides Theme"**.

You will get a copy of this document on your Google Drive and will be able to edit, add or delete slides.

You have to be signed in to your Google account.

EDIT IN POWERPOINT®

Click on the button under the presentation preview that says **"Download as PowerPoint template"**. You will get a .pptx file that you can edit in PowerPoint.

Remember to download and install the fonts used in this presentation (you'll find the links to the font files needed in the Presentation design slide)

Hello!

I am Jayden Smith.

I am here because I love to give presentations. You can find me at @username

1.
Transition headline

Let’s start with the first set of slides

“ Quotations are commonly printed as a means of inspiration and to invoke philosophical thoughts from the reader.

This is a slide title

- Here you have a list of items
- And some text
- But remember not to overload your slides with content

Your audience will listen to you or read the content, but won't do both.

Big concept

Bring the attention of your audience over a key concept using icons or illustrations

Business Conclusions

You can also split your content

White

Is the color of milk and fresh snow, the color produced by the combination of all the colors of the visible spectrum.

Black

Is the color of coal, ebony, and of outer space. It is the darkest color, the result of the absence of or complete absorption of light.

RFE

R^2 **0.76**

BASELINE

57 Features

R^2 **0.87**

PEAKS DETECTION

58 Features

RFE

R^2 **0.86**

54 Features

R^2 **0.83**

HOUR BINS

40 Features

RFE

R^2 **0.82**

36 Features

4 Features Eliminated:

Humidity | Actual Temperature | Wind Speed | Working Day

Maps

Optimization Using Data

Maintenance & Repair:

Data driven approach to optimize processes to keep bikes and docks in good repair, safe, and available.

Technologies:

Usage of geofencing

Future Usage Modeling

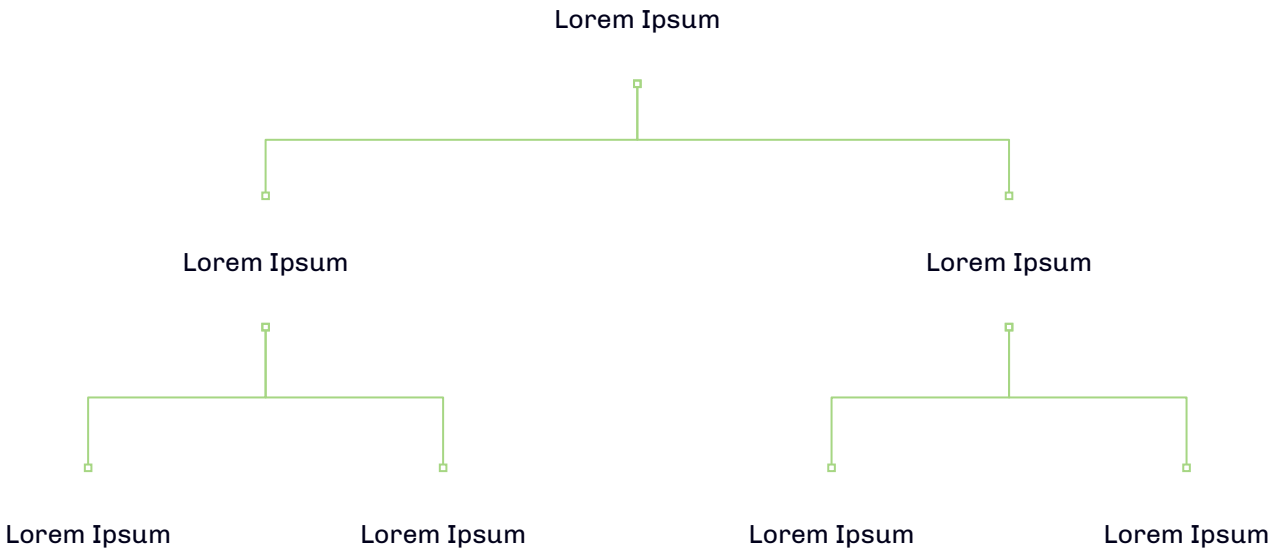
Is the color of blood, and because of this it has historically been associated with sacrifice, danger and courage.

A picture is worth a thousand words

A complex idea can be conveyed with just a single still image, namely making it possible to absorb large amounts of data quickly.

Want big impact?
Use big image.

Use diagrams to explain your ideas



And tables to compare data

	A	B	C
Yellow	10	20	7
Blue	30	15	10
Orange	5	24	16

89,526,124

Whoa! That’s a big number, aren’t you proud?

89,526,124$

That’s a lot of money

185,244 users

And a lot of users

100%

Total success!

Let’s review some concepts

Yellow
Is the color of gold, butter and ripe lemons. In the spectrum of visible light, yellow is found between green and orange.

Blue
Is the colour of the clear sky and the deep sea. It is located between violet and green on the optical spectrum.

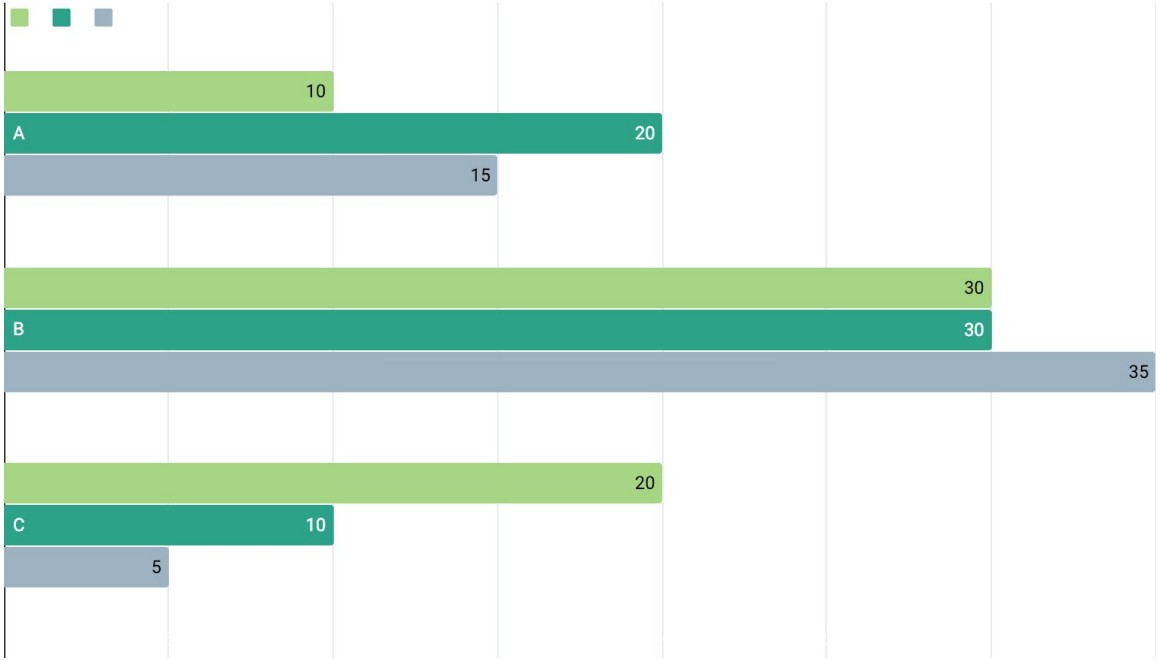
Red
Is the color of blood, and because of this it has historically been associated with sacrifice, danger and courage.

Yellow
Is the color of gold, butter and ripe lemons. In the spectrum of visible light, yellow is found between green and orange.

Blue
Is the colour of the clear sky and the deep sea. It is located between violet and green on the optical spectrum.

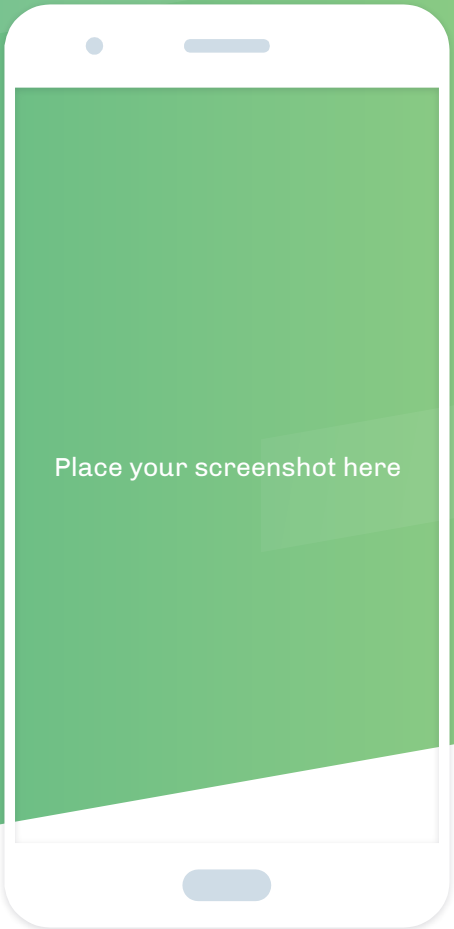
Red
Is the color of blood, and because of this it has historically been associated with sacrifice, danger and courage.

You can insert graphs from Google Sheets



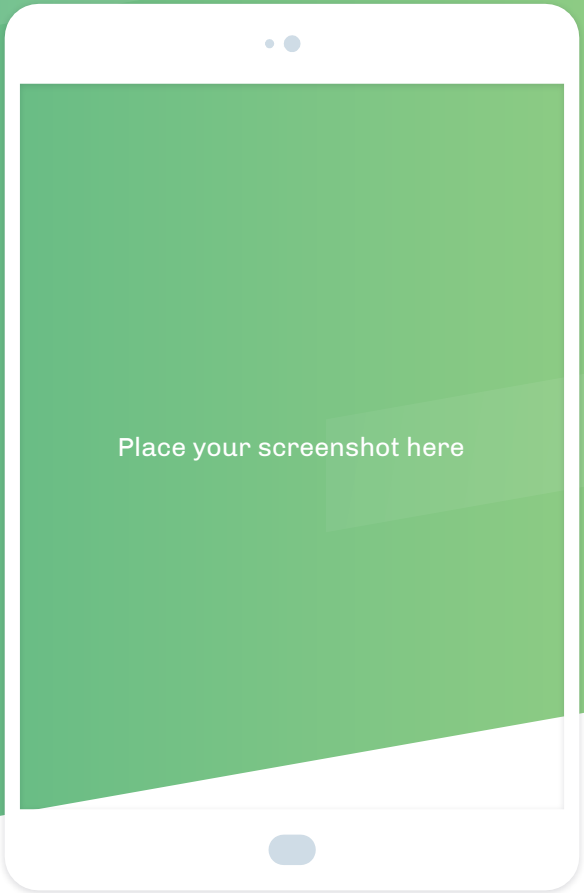
Mobile project

Show and explain your web, app or software projects using these gadget templates.



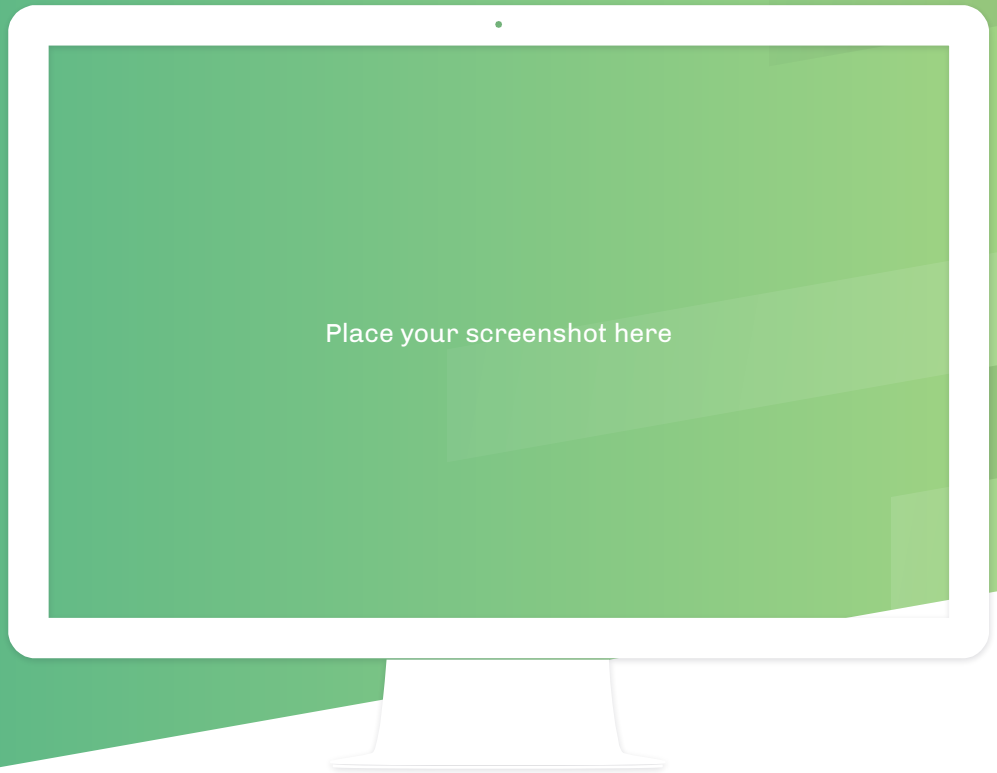
Tablet project

Show and explain your web, app or software projects using these gadget templates.



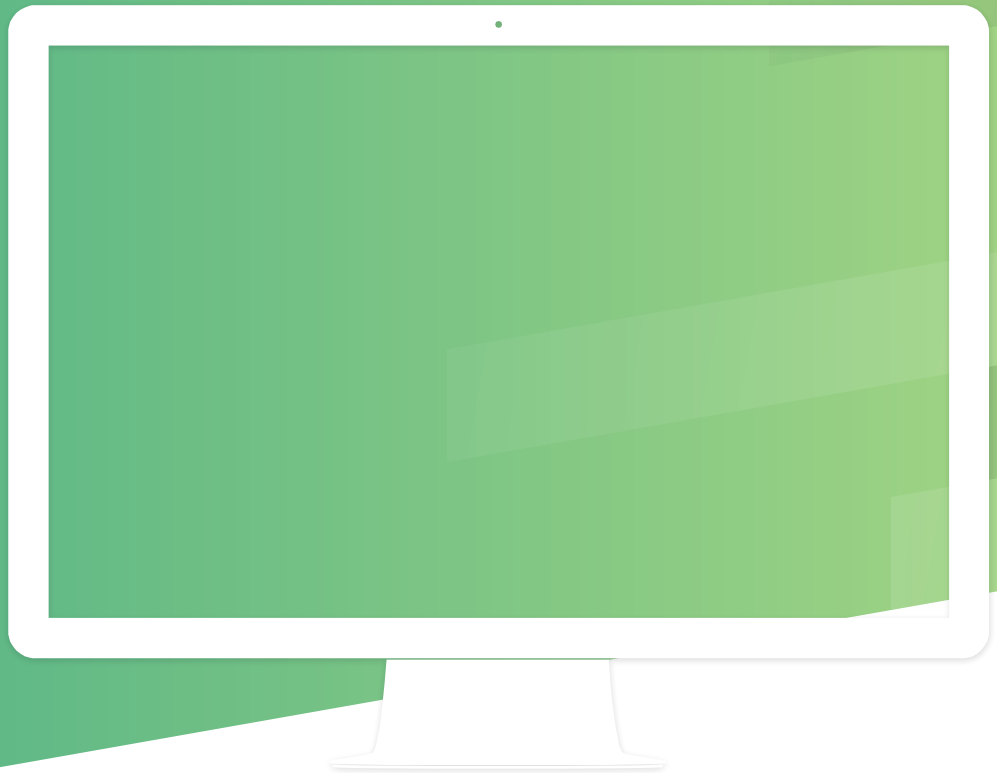
Desktop project

Show and explain your web, app or software projects using these gadget templates.



Desktop project

Show and explain your web, app or software projects using these gadget templates.



Thanks!

Any questions?

You can find me at:

- @username
- user@mail.me

Credits

Special thanks to all the people who made and released these awesome resources for free:

- Presentation template by SlidesCarnival
- Photographs by Unsplash

Presentation design

This presentation uses the following typographies and colors:

- Titles: Roboto Slab
- Body copy: Chivo

You can download the fonts on these pages:

https://www.fontsquirrel.com/fonts/roboto-slab

https://www.fontsquirrel.com/fonts/chivo

Lime **#a6d683** / Clover **#2ca388** / Fog **#9eb3c2**

You don't need to keep this slide in your presentation. It's only here to serve you as a design guide if you need to create new slides or download the fonts to edit the presentation in PowerPoint®

SlidesCarnival icons are editable shapes.

This means that you can:

- Resize them without losing quality.
- Change fill color and opacity.
- Change line color, width and style.

Isn't that nice? :)

Examples: