* [Overview](#gjdgxs)
* [Tips from Beatriz](#30j0zll)

# CSE 450 - Machine Learning & Data Mining

## Module 03 — Targeted Marketing, Project

## Overview

After a few more meetings, Beatriz has assigned your team to address the following issues asked by the stakeholders:

##### 

Miguel Ferreira, Bank President asks:

Like I said the other day, the core task we're interested in is identifying those customers most likely to subscribe to a term deposit.

This way, we can build a targeted marketing campaign that focuses primarily on those customers.

##### 

Francisco, VP of Marketing says:

And I'd like you to find any actionable patterns in our results. Should we only call single people on Saturdays? Does it make sense to call students at all?

Things like that.

##### 

Miguel Ferreira, Bank President adds:

One other thing we should probably address, does contacting people too frequently for these marketing campaigns have an adverse affect on the outcome?

##### 

Beatriz, Senior Data Scientist says:

One last thing, there are a bunch of social and economic indicators in the data.

We should be careful about how we consider these. We may want to see separate models for times when, for example, the consumer confidence index is high compared to when it is low.

##### 

Miguel Ferreira, Bank President adds:

Good thought Beatriz. Different customer segments tend to react to economic changes differently.

We'll definitely want to know which model is best to use during different economic situations.

##### 

Beatriz, Senior Data Scientist says:

One last thing, since we're planning on deploying these models, we'll want to make sure that once you have the models trained and tested, that you [persist those trained models to a file](https://scikit-learn.org/stable/modules/model_persistence.html) so we can load them into our production systems.

##### 

Miguel Ferreira, Bank President adds:

Sounds like we have enough to get started. If you could send us [your write up on this by Saturday night](http://docs.google.com/summary.docx), that would be great.

##### 

Beatriz, Senior Data Scientist says:

Here are some resources I've put together to help your team out on this project.

## Tips from Beatriz

Data Dictionary

Our database analyst put together this [data dictionary](http://docs.google.com/bank-dictionary.txt) to help explain the values and sources of different columns in the [bank dataset](https://raw.githubusercontent.com/byui-cse/cse450-course/master/data/bank.csv), so be sure to review that.

Target Variable

One oddity here is that our target feature is simply labled y, but it's a boolean indicating "y" or "n", did the client subscribe to a term deposit.

Social and Economic Context

All of the features listed as "social and economic context attributes" should be treated with care. Since these attributes are associated with the broader national condition rather than specific customer attributes, we'll want to treat these features differently.

Creating different models for "good times" and "hard times" might be a good start.

Decision Trees

You can find documentation on how to use decision trees with sci-kit learn on these pages:

* [User Guide Entry for Decision Trees](https://scikit-learn.org/stable/modules/tree.html)
* [API Reference for DecisionTreeClassifier](https://scikit-learn.org/stable/modules/generated/sklearn.tree.DecisionTreeClassifier.html)
* [Tips for Practical Use, from the User Guide](https://scikit-learn.org/stable/modules/tree.html#tips-on-practical-use)

Model Persistence

When you train a model, a large amount of information is stored in memory. That model can then be used to make predictions for new instances at a later time.

You'll want to save these trained models using python's pickle module, [as shown here](https://scikit-learn.org/stable/modules/model_persistence.html).

However, rather than using pickle's default protocol version, you should use [protocol version 5](https://docs.python.org/3/library/pickle.html#data-stream-format), which was introduced in Python 3.8 and is optimized for dealing with structures that contain numpy arrays and pandas data frames.

Model Ensembles, Bagging, and Boosting

Often, we can get better results by using a set of models, each using a slightly different set of training data, or other parameters. These are called "Model Ensembles" and it's very common to use an ensemble of decision trees (often called a "Random Forest") rather than a single tree.

Two popular techniques used in the creation of ensembles are "boosting" and "bagging". You can read more about these topics on pages 163 - 167 of your textbook.

For details on how to use these techniques with Sci-Kit Learn, see [this page](https://scikit-learn.org/stable/modules/ensemble.html).

Avoiding overfitting through pruning

It is *very* easy to overfit a decision tree. The text discusses strategies to avoid this problem in section 4.4.4 (pages 158 - 163).

In SciKit-Learn, you can use [parameters such as max\_depth and min\_samples\_leaf](https://scikit-learn.org/stable/modules/generated/sklearn.tree.DecisionTreeClassifier.html) to control tree complexity and overfitting.

Alternatively, you can use something more elaborate, such as [cost complexity pruning](https://scikit-learn.org/stable/auto_examples/tree/plot_cost_complexity_pruning.html#sphx-glr-auto-examples-tree-plot-cost-complexity-pruning-py).

##### 

Johnny, the Data Science Intern, sends you an email:

Hey, I heard you were working here in Portugal, that's so cool! Beatriz is actually my cousin, and she filled me in on some of the project details. I hope you don't mind, but I put together [this collection of Google Colab notebooks](http://docs.google.com/hints.html) that might be useful.

1. [CEO photo by Oz Seyrek on Unsplash](https://unsplash.com/photos/-Ir03_pgpMU)  [↩](#1fob9te)
2. [VP of HR photo by Christina @ wocintechchat.com](https://unsplash.com/photos/SJvDxw0azqw) [↩](#3znysh7)
3. [VP of Finance photo by steffen Wienberg on Unsplash](https://unsplash.com/photos/ml-pxK0Ovmw) [↩](#2et92p0)
4. [Data Science Intern photo by Fábio Lucas on Unsplash](https://unsplash.com/photos/iczrMDNuvzkml-pxK0Ovmw) [↩](#tyjcwt)
5. [Data Science Intern photo by Fábio Lucas on Unsplash](https://unsplash.com/photos/iczrMDNuvzkml-pxK0Ovmw) [↩](#3dy6vkm)