Project Proposal

Packt GP 1

1. Title

Customer prioritization for the sales team.

2. Team

2.1 Number of participants

Minimum: 2 (1 data scientist and 1 app developer)
Recommended: 4 to 6 (data scientists and app developers)

Finalised: 4 (3 data scientists and 1 app developer)

2.2 Profile

Role in Project	User Story	Minimum Skillset	Candidate Job Titles	Number of personnel required	Mandatory role?
Project Lead				1	Yes
Owner / Sponsor	As a sponsor, I want to ensure adequate resources and support is available to the team, so that we can produce good ROI.		Business Head, COO	1+	Yes
Data Science Team Lead	As a data science team lead, I want to ensure the model we build produces at least 50% more conversions than the baseline, so that we can make more sales and thus more profit.		Lead Data Scientist	1	Smaller budgets usually combine this role with another.
App Development Team Lead / UX Lead	As an app dev team lead, I want to ensure we build a simple yet robust interface for the model, so that our sales executives can get more done in less time.		Lead Developer, UX Lead	1	Smaller budgets usually combine this role with another.
Testing Team Lead	As a testing team lead, I want to ensure the data product we build passes stress tests, in addition to checking if all the functional requirements are satisfied as well, so that we can deliver a functional and robust product.		Testing Team Lead	1	Yes, although for learning purposes, we will skip this role.
Data Scientist	As a data scientist, I want to analyze the data thoroughly, set up a preprocessing pipeline and experiment with candidate models, so that we can settle on a final model that satisfies the solution requirement.	a) +2 level Science, Math, English b) Proficient in using Python c) Proficient in Python DS libraries, including Numpy, Pandas, Seaborn and Scikit-learn	Data Scientist	1+	Yes
Data Engineer	As a data engineer, I want to ensure that the right data is available at the right time to the data scientists, so that they can analyze it and use it to create the model.	a) Proficient in SQL and NoSQL Databases b) Proficient in data ingestion tools and techniques	Data Engineer	1+	Smaller budgets usually combine this role with another.
App Developer	As an app developer, I want to analyze the input output requirements of the model and develop an interface that provides easy access to our sales executives with an always up-to-date list of customers sorted by priority, so that more potential customers can be catered to in a given amount of time.	a) Proficient in any web app development stack, such as HTML/CSS/Javascript/PHP	App Developer, Web Developer	1+	Yes
Test Engineer	As a test engineer, I want to create a customized process to execute all the required tests for the data product, so that the users can confidently utilize it.		Test Engineer	1+	Yes, although for learning purposes, we will skip this role.

3. Timeline

3.1 Duration

10 working days (2 calendar weeks)

3.2 Planned Start Date

10th September, 2021

3.3 Planned End Date

23rd September, 2021

4. Technology Stack

Data Science: Python, Pandas, Seaborn, Scikit-learn
Web App : Any (example: HTML/CSS/Javascript/PHP)

5. Resources

5.1 Recommended Preparation and Study Material

Topics:

Python Design Aphorisms
https://www.python.org/dev/peps/pep-0020/

Python Coding Conventions
https://www.python.org/dev/peps/pep-0008/

Python Syntax and Semantics
https://docs.python.org/3/reference/index.html
https://docs.python.org/3/library/index.html

Python OOP
https://docs.microsoft.com/en-us/learn/modules/python-object-oriented-programming/
https://realpython.com/python3-object-oriented-programming/
https://www.educative.io/blog/object-oriented-programming
https://www.educative.io/blog/how-to-use-oop-in-python
https://docs.python.org/3/tutorial/classes.html

EDA
https://arxiv.org/pdf/2104.00841.pdf
https://datascience.foundation/sciencewhitepaper/data-analysis-with-pandas
https://hackernoon.com/best-libraries-that-will-assist-you-in-eda-2021-edition-2l1734hl
https://www.packtpub.com/product/hands-on-exploratory-data-analysis-with-python/9781789537253
https://ai.plainenglish.io/exploratory-data-analysis-eda-with-python-matplotlib-bb784e1d3dd3
https://towardsai.net/p/data-analysis/exploratory-data-analysis-in-python-ebdf643a33f6

Data Wrangling / Preparation
https://towardsdatascience.com/essential-commands-for-data-preparation-with-pandas-ed01579cf214
https://www.explorium.ai/blog/top-tips-for-data-preparation-using-python/
https://realpython.com/python-data-cleaning-numpy-pandas/
https://data-flair.training/blogs/data-wrangling-with-python/
https://devopedia.org/data-preparation

Machine Learning
https://www.sciencedirect.com/topics/computer-science/machine-learning
https://www.oracle.com/in/data-science/machine-learning/what-is-machine-learning/
https://scikit-learn.org/stable/auto_examples/classification/plot_classifier_comparison.html
https://scikit-learn.org/stable/modules/generated/sklearn.linear_model.LogisticRegression.html
https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.RandomForestClassifier.html
https://scikit-learn.org/stable/modules/generated/sklearn.linear_model.SGDClassifier.html

Additional Resources (Free):

https://courses.packtpub.com/courses/python
https://docs.python.org/3.7/
https://numpy.org/
https://pandas.pydata.org/docs/
https://seaborn.pydata.org/
https://scikit-learn.org/stable/

Yet More Resources (Paid):

https://courses.packtpub.com/courses/data-analysis
https://courses.packtpub.com/courses/data-wrangling
https://courses.packtpub.com/courses/supervised-learning
https://courses.packtpub.com/courses/machine-learning
https://courses.packtpub.com/courses/data-science

5.2 GitHub Repo

https://github.com/TeamEpicProjects/Customer-Prioritization-for-Marketing

5.3 Datasets

6. Recommended System Setup

6.1 Hardware Requirements

4-core CPU with 8GB RAM and 40GB drive space.

6.2 Software Requirements

PyCharm 2021+
- python 3.6+
  - pandas 1.1+
  - pandarallel 1.5+
  - seaborn 1.1+
  - sklearn 0.23+
  - sqlalchemy 1.3+
  - pymysql 0.9+
  - joblib 0.15+

7. Scope

7.1 Problem Statement

Thousands of potential customers visit our website every day for a free horoscope report, some of which actually result in a conversion.

Due to limited human resources, we are unable to reach out to each one of those thousands of potential customers each day. In addition to being infeasible for us, it is probably not necessary either.

To help filter the long list into something manageable by the sales team, we came up with a baseline model that prioritizes the customers we reach out to each day. The baseline model was built in a hurry without any serious data analysis and it is just a static formula taking as input certain values generated from browsing sessions.

Over the years, we have found that a majority of the potential customers we reach out to do not result in an immediate conversion.

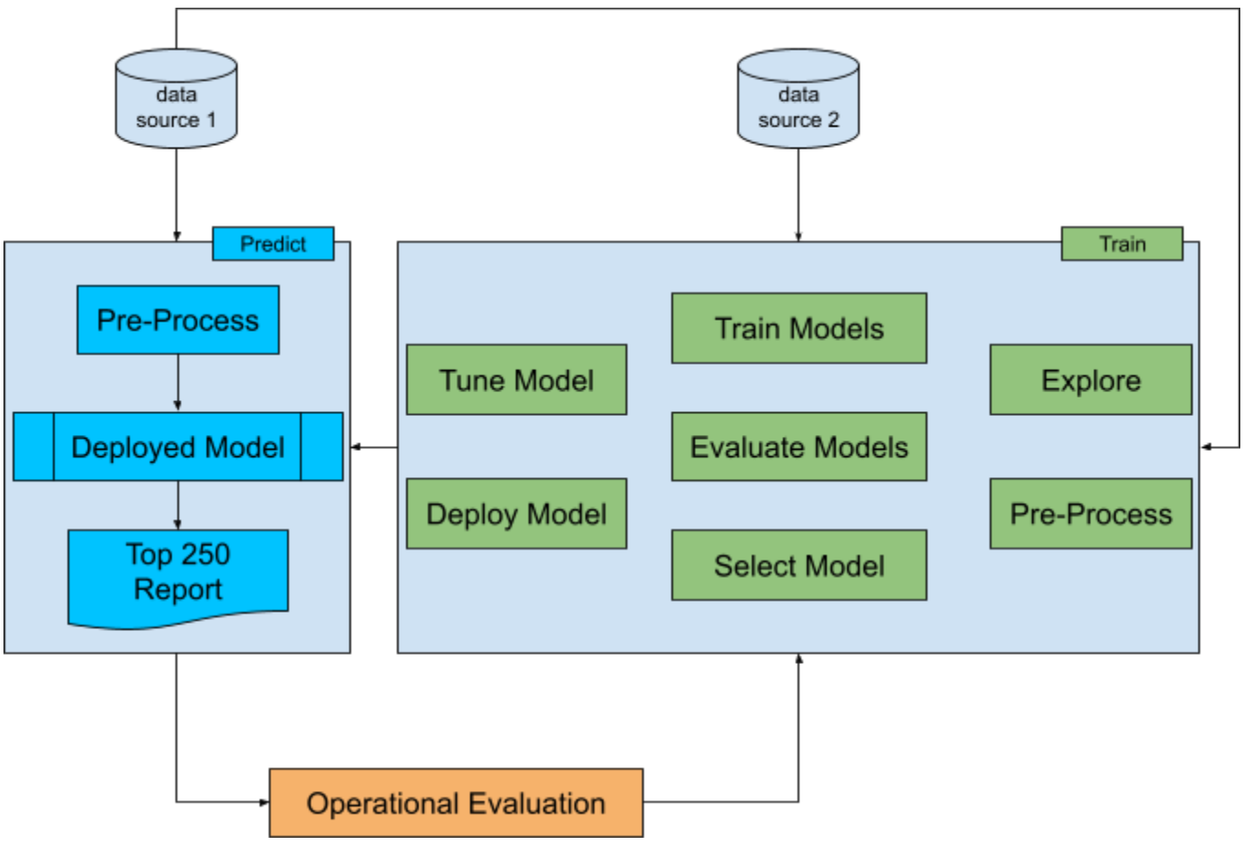
We want to use data and technology to maximize conversions from our contacts each day.

7.2 Expected Solution

The solution should consist of the following:

a) A report containing key summaries and insights from historical data.
b) A data science product using an ML model to prioritize the potential list of customers.
   The predictions report (filename format: top_250_report_ddmmyyyy.csv) should be limited to the top 250 customers with the following fields:
   customer_id, conversion_probability
   The model needs to consider only those customers who visited our website within the last 48 hours.
c) An evaluation report showing at least 50% increase in the conversions when compared with the baseline model (over a period of 7 days).

7.3 HLA Diagram



7.4 Evaluation Criteria

Each project participant would be assessed and awarded a score based on the following criteria:

a) Fulfillment of expected solution: Team achievement score, 30% weight
   i) Accuracy (Balanced Accuracy): 80% weight
   ii) Function: Output needs to meet specification, given an input set: 20% weight

b) Individual contribution / performance: 70% weight
   i) Completion of daily tasks, quality of completed tasks: 80% weight
      - Completing tasks on schedule: 20%
      - Code: 30%
        (a) Quality: 60%
        (b) Correctness: 40%
      - Documentation: 30%
        (a) Completeness: 50%
        (b) Quality: 50%
      - Presentation: 20%
   ii) Lateral thinking / extra mile: 20% weight

7.5 Work Package

Day	Milestones	Deliverables	Individual Tasks
1	a) Understanding the problem and the expected solution b) Setting up hardware and software c) High-level understanding of the data (input) d) Exploring ways to solve the problem	a) Platform ready report b) Data summary report (basic) c) Solution candidates report	
2	a) HLA / solution design b) EDA Level 1	a) Solution design document b) Data insights and summary report (detailed)	
3	a) Data preprocessing b) Feature engineering / selection	a) Report on prepared data	
4	a) Creating a non-ML baseline model b) Creating and testing the pipeline	a) Solution report	
5	a) Wrapping a product in a user-friendly interface	a) Solution report	
6	a) Creating and evaluating base models	a) Candidate models report (base)	
7			
8	a) Tuning the models	a) Candidate models report (tuned)	
9			
10	a) Model selection and fine tunings b) Final testing and evaluation c) Presenting the solution	a) Solution report	

Please note that each day, the group's task would be to review previous day's deliverables.
Individual tasks would be assigned during onboarding meet.