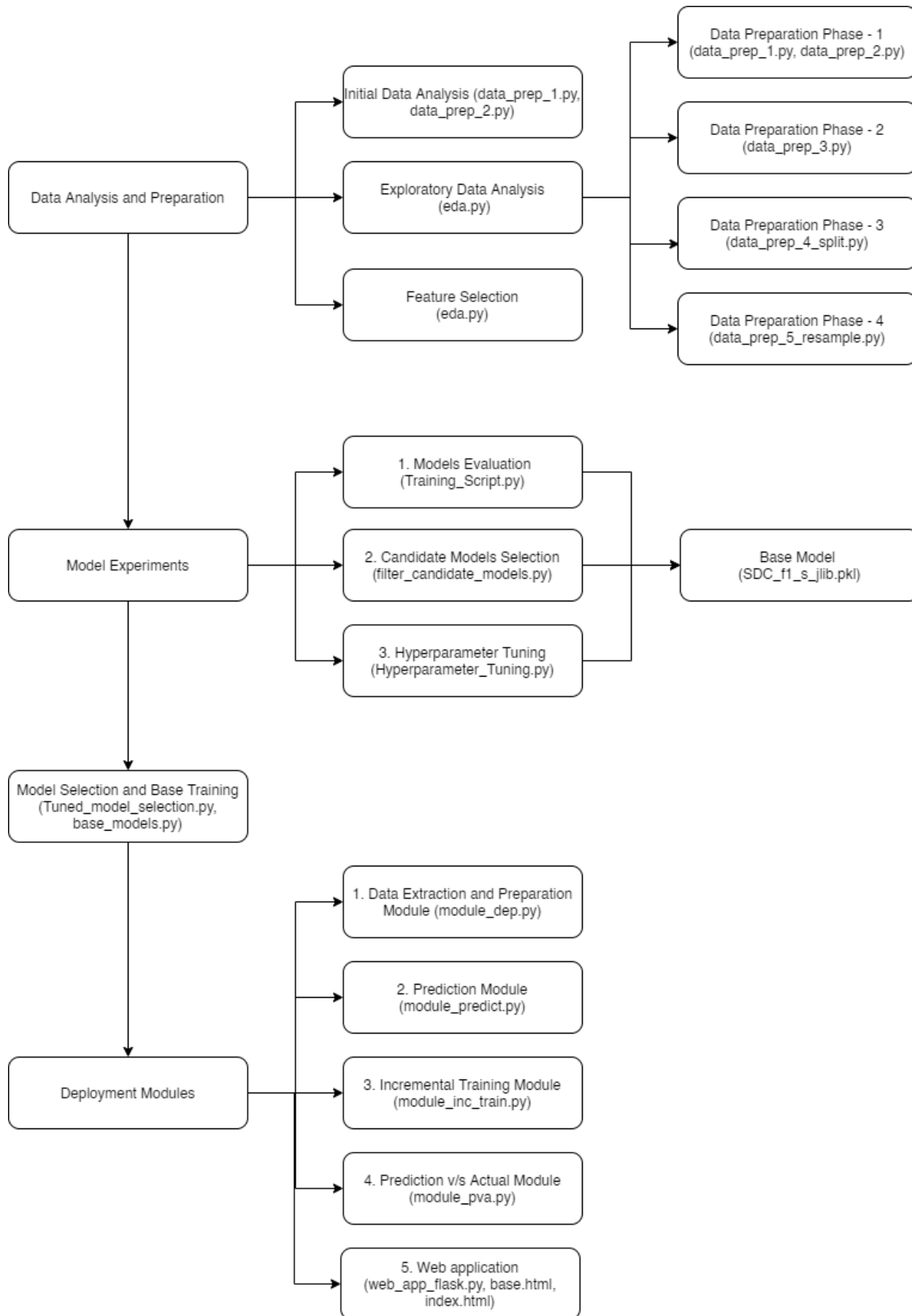


Packt Group Project: Customer Prioritization for Sales Solution Development Summary

Development	2
Flowchart	2
Summary	3
Data Analysis and Preparation	3
Model Experiments	3
Deployment	3
Infrastructure used	4
Hardware	4
Software	4
Additional Notes	5
Deployment	5

Development

Flowchart



Summary

Data Analysis and Preparation

- There were 6 datasets available for our solution development, only 5 were used.
- They were all consolidated and merged to create a final base data set.
- The dataset was split into two for development and operational purposes.
- Exploratory data analysis was performed to check the count plots, detect the outliers and create a correlation heatmap.
- Due to the imbalanced nature of the target variable, Tomek resampling was used to balance it.
- Exploratory data analysis was performed once again to check for the presence of outliers and to create a new correlation heatmap.
- From the final correlation heatmap, features were selected for experimentation with the Machine Learning models.

Model Experiments

- Training Script contains all the Machine Learning algorithms we used to experiment with our dataset. It also contains the experiments with all the different feature sets that we chose from the final correlation table.
- Based on all the models that we obtained from the Training script, filter candidate model script was used to filter models based on a Balanced Accuracy and Recall threshold set by the Project Lead.
- After making a selection of models, we tuned their hyperparameters in the Hyperparameter Tuning script.
- Tuned model selection script cross validated the tuned models to obtain a final table with modified Balanced Accuracy and Recall scores.
- Base models script contains the final choice of top models that we obtained after tuned cross validation.
- SDC_f1_s_jlib: Stochastic Gradient Descent Classifier with feature set 1 standard scaled was chosen as the final model for deployment.

Deployment

- Module dep does two things: It converts the date column into datetime format and filters the data for the requested date.
- Module prediction makes prediction using SDC_f1_s_jlib.pkl model.
- Module inc train retrains the machine learning model on the new data received.
- Module pva generates a report of predicted vs actual target variable.
- Web application contains the script that creates the web app interface with all the modules connected to return the requested report for a certain date.

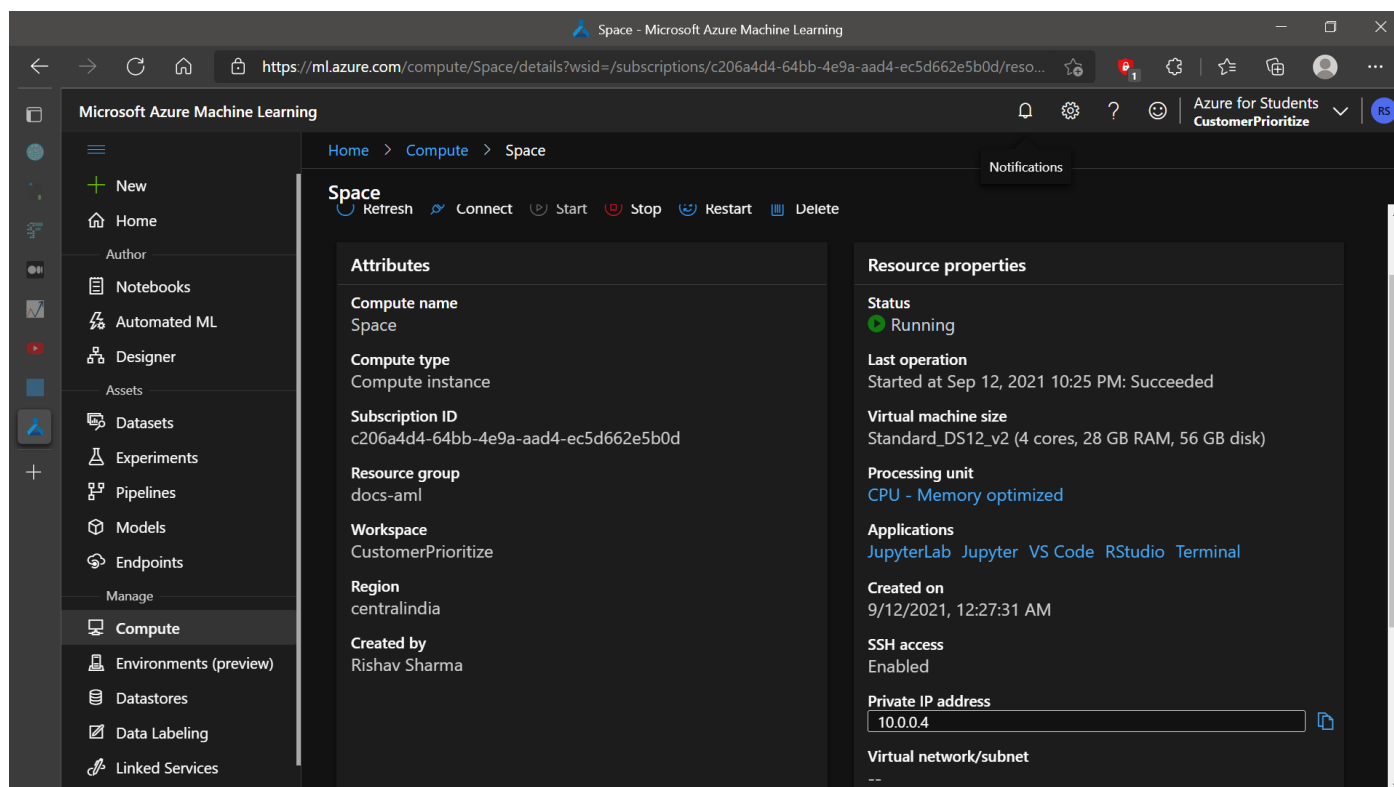
Infrastructure used

Hardware

Local Machine:

Processor	Intel Core i3
Available Disk Space	300 GB
RAM	4 GB

Cloud Machine:



The screenshot displays the Microsoft Azure Machine Learning web interface. The left sidebar contains navigation options: Home, Author (Notebooks, Automated ML, Designer), Assets (Datasets, Experiments, Pipelines, Models, Endpoints), and Manage (Compute, Environments (preview), Datastores, Data Labeling, Linked Services). The main panel shows the 'Space' compute instance details. The breadcrumb navigation is 'Home > Compute > Space'. The instance is in a 'Running' status. Key attributes include: Compute name: Space, Compute type: Compute instance, Subscription ID: c206a4d4-64bb-4e9a-aad4-ec5d662e5b0d, Resource group: docs-aml, Workspace: CustomerPrioritize, Region: centralindia, and Created by: Rishav Sharma. The resource properties section shows: Status: Running, Last operation: Started at Sep 12, 2021 10:25 PM: Succeeded, Virtual machine size: Standard_DS12_v2 (4 cores, 28 GB RAM, 56 GB disk), Processing unit: CPU - Memory optimized, Applications: JupyterLab, Jupyter, VS Code, RStudio, Terminal, Created on: 9/12/2021, 12:27:31 AM, SSH access: Enabled, Private IP address: 10.0.0.4, and Virtual network/subnet: --.

Space - Microsoft Azure Machine Learning

https://ml.azure.com/compute/Space/details?wsid=/subscriptions/c206a4d4-64bb-4e9a-aad4-ec5d662e5b0d/reso...

Microsoft Azure Machine Learning

Home > Compute > Space

Space

Refresh Connect Start Stop Restart Delete

Attributes

Compute name
Space

Compute type
Compute instance

Subscription ID
c206a4d4-64bb-4e9a-aad4-ec5d662e5b0d

Resource group
docs-aml

Workspace
CustomerPrioritize

Region
centralindia

Created by
Rishav Sharma

Resource properties

Status
Running

Last operation
Started at Sep 12, 2021 10:25 PM: Succeeded

Virtual machine size
Standard_DS12_v2 (4 cores, 28 GB RAM, 56 GB disk)

Processing unit
CPU - Memory optimized

Applications
JupyterLab Jupyter VS Code RStudio Terminal

Created on
9/12/2021, 12:27:31 AM

SSH access
Enabled

Private IP address
10.0.0.4

Virtual network/subnet
--

Software

- Anaconda Distribution
- Spyder IDE
- Python – numpy, pandas, pandarallel, matplotlib, seaborn, scikit-learn, joblib, imblearn.

Additional Notes

- a) Note from the team lead: The datasets contain real data from a real-world problem, with certain content masked (sanitised) in order to preserve privacy and other laws. For example customer emails and phone numbers are masked and represented as numeric values - this does not affect the problem or its solution in any way.
- b) Although the expected solution involved creating a non-ML base model for benchmarking purposes, we decided to skip it (after consulting team lead) due to lack of resources.

Deployment

We deployed our solution for demo purposes @ <https://packt-cpm.herokuapp.com/>
This required two additional files:

- a) requirements.txt
- b) Procfile

We followed the guide @ <https://stackabuse.com/deploying-a-flask-application-to-heroku/>

Find the solution @ Github repo <https://github.com/TeamEpicProjects/Customer-Prioritization-for-Marketing>