**Capstone Project**

**Document Skeleton**

# Process overview

The following diagram shows the overall end-to-end process for defining, designing and delivering the Capstone project.

Diagram

Description automatically generated

Note: The following are the candidate sections of the document. They are presented here for guidance. Questions in each section could be used as possible aspects to cover. Some questions may not be applied to each project. On the other hand, additional information may be needed.

# Problem statement

* What is the problem or the opportunity that the project is investigating?
* Improve the e-commerce company customer conversion rate with recommender system implementation and optimization
* Why is this problem valuable to address?
* Improve customer retention and engagement
* Increase e-commerce company sale revenue
* Understand the market insight for better informed business decision
* What is the current state (e.g. unsatisfied customers, lost revenue)?
* Almost all the e-commerce companies use recommender system to help user discover new and relevant items
* 40-70% of the big companies’ revenue comes directly from recommender system
* What is the desired state?
* Improve the target company customer conversion rate to above 70%
* Has this problem been addressed by other research projects? What were the outcomes?
* Different companies implement different models and algorithms, and the results are improving but not to the target level for whole industry.

# Industry/ domain

* What is the industry/ domain?
* E-commerce industry
* What is the current state of this industry?
* Same as mentioned in problem statement
* What is the overall industry value-chain?
* E-commerce helps companies reduce overall costs, improve data accuracy, streamline supply chain services, accelerate business cycles, and enhance customer service.
* What are the key concepts in the industry?
* The key concepts in Ecommerce industry include personalization, product recommend, cart reminder, product social share, product review and response design
* Is the project relevant to other industries?
* Yes, the project is relevant to search engine and social media

# Stakeholders

* Who are the stakeholders?
* Ecommerce company like Netflix or Youtube management team
* Why do they care about this problem?
* This project can help improve the customer conversion with recommender system
* What are the stakeholders’ expectations?
* Beat the current baseline and improve as much as possible

# Business question

* What is the main business question that needs to be answered?
* How can we improve customer conversion rate?
* What is the business value of answering this question?
* The sale revenue will increase $250M if 1% customer conversion is improved for Netflix
* The sale revenue will increase $200M if 1% customer conversion is improved for Youtube

# Data question

* What is the data question that needs to be answered?
* How can we recommend the items customers will most possibly like?
* What is the data required to answer the question?
* The historic data for people’s interaction with products or service

# Data

* Where was the data sourced?
* MovieLens 100K from Kaggle website
* What is the volume and attributes of the data?
* Total 100,004 movie ratings across 671 users and 9066 movies
* How reliable is the data?
* The data sets were collected by the GroupLens Research Project at the University of Minnesota and it is used for the Kaggle competition.
* What is the quality of the raw data? How was this data generated?
* The data was collected through the MovieLens web site  
  (movielens.umn.edu) during the seven-month period from September 19th,  
  1997 through April 22nd, 1998. This data has been cleaned up - users  
  who had less than 20 ratings or did not have complete demographic  
  information were removed from this data set.
* Is this data available on an ongoing basis?
* Yes.

# Data science process

## Data analysis

* What data pipeline was to wrangle the raw data?
* Data extraction->Data Process->Model Training->Model Validation
* What are the highlights of the Exploratory Data Analysis (EDA)? NA
* Is the pipeline reusable? YES
* What are the intermediary data structures used (if any)? NA

## Modelling

* What are the main features used? UserID & MovieID
* Did you find any interesting interactions between features? NA
* Is there a subset of features that would get a significant portion of your final performance? Which features? NA
* How did you select features? NA
* What feature engineering techniques are used? NA
* What are the models used?
* Best Seller recommendation
* Matrix Factorization
* Neural Collaborative Filtering
* How long does it take to train your model? 3 hours
* What are the tools used? Jupiter Notebook
* What are the model performance metrics? MSE & Hit Rate
* Which model was selected? Neural Collaborative Filtering

## Outcomes

* What are the main findings and conclusions of the data science process?
* MSE (Mean Square Error) scoring is 0.002 – 0.003
* Hit Rate @ 10 of best seller is 5-10%
* Hit Rate @ 10 of Matrix Factorization is 60-70%
* Hit Rate @ 10 of Neural Collaborative Filtering is 70-80%

## Implementation

* What are the considerations for implementing the model in production?
* Add best seller to solve cold start problem initially
* Incorporate user’s profile to solve cold start problem at later stage

# Data answer

* Was the data question answered satisfactorily? Yes, with the selected model
* What is the confidence level in the data answer? High, based on the hit rate score for model valuation

# Business answer

* Was the business question answered satisfactorily? Yes
* What is the confidence level in the business answer? High, based on model valuation

# Response to stakeholders

* What are the overall messages and recommendations to the stakeholders?
* Hit rate of recommender system is critical to customer conversion rate
* The model hit rate outperforms current baseline 60% for Netflix
* Deploy the model into the recommender system engine
* Explore user new interests with Multi-armed bandit reinforcement learning method

# End-to-end solution

* What is the overall end-to-end solution to use the model developed in the project?

Diagram

Description automatically generated

# References

* Where are the data and code used in the project?
* Dataset of MovieLens 100K from Kaggle
  + https://www.kaggle.com/rajmehra03/movielens100k
* Project Github
  + https://github.com/dyespring/Institute\_of\_data/tree/main/Capstone\_Project
* What are the resources used in the project?
* Neural Collaborative Filter Github
  + https://github.com/microsoft/recommenders