ML Engineer Code Test
Pradeep Medagiri [130 mins]

I. Task 1 [30mins]

II. Task 2 [40 mins]

III. Task 3 [60 mins]

1. Initial Exploratory Data Analysis
2. Data Preprocessing
3. Feature Engineering
4. Model Development
5. Model Evaluation

Task 1

- Performed in Jupyter notebook, file name: “Task1&2.ipynb”

Sample output:

```
[{'date': datetime.datetime(2012, 1, 1, 0, 0), 'Page Views': '0', 'Visi
ts': '0', 'Unique Visitors': '0', 'Bounce Rate': 'INF'},
```

Task 2

- Performed in Jupyter notebook, file name: “Task1&2.ipynb”

Sample output:

```
[{'GeoSegmentation Country': 'United States', 'GeoSegmentation City': '
los angeles (California, United States)', 'software title (v64)': '::un
specified::', 'page_views': '4418', 'visits': '1924'},
```

Task 3

Raw data:

Table 1: Subscription Details – 8 columns
Table 2: Video Consumption Details – 13 columns
Table 3: Profile Details – 3 columns
Table 4: Demographic Details – 3 columns

Data Integration:
Import all four tables and read as pandas data frame in Jupyter notebook. Combine all tables using merge or joins on "Unique Registration ID".

Total columns: 24

1. Initial Exploratory Data Analysis: Visualize data using below libraries in python or using Tableau.

Data visualization:
Tableau
Seaborn
Matplotlib
ggplot

- First, plot a bar graph to visualize percentage of conversions "*converted_ind*" (To check data distribution)
- Plot a Correlation matrix that provide information on correlated columns.
- Plot for Univariate, Multivariate Analysis.
- Insights drawn from above steps after real data is plotted and this helps to continue process in data preprocessing.
For example: missing data, mislabeled data, grouping, clustering required (yes or no).

2. Data Preprocessing

Python, TensorFlow, Jupyter notebook, google colab, MLflow

i. Less correlated columns are removed.
ii. Check for missing/nan values.
- Drop rows using pandas
- Imputation techniques (box plot above)

Perform above actions depending on business problem and business data.

Check data types – dataframe.dtypes

1. Replace values: Boolean type [0 if False, 1 if True]
Columns: "*converted_ind,*
video_25_pct_ind,

video_75_pct_ind,
video_content_complete_ind"

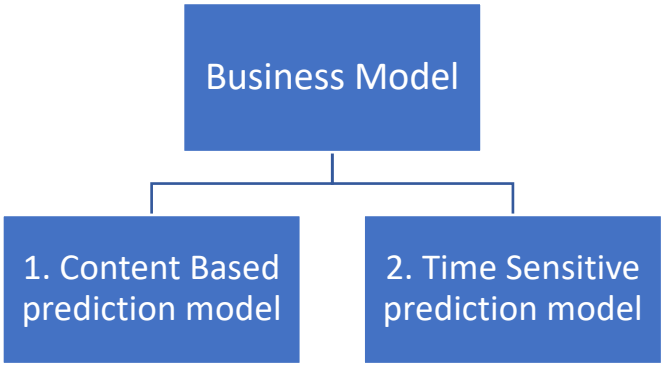
2. Categorical columns having 2 or more labels, implement one-hot-encoding.
3. Depending on business model, we have to consider below metrics:
    1. Conversion score/rate "*paid_conversion_score*"
    2. Abandon rate
    3. Churn rate
    4. Engagement on platform
    5. Active/Inactive

Target variable:
Option 1. "converted_ind" as target variable.
Option 2. From above metrics, create a target variable and mark boundaries/range that classify's as **converted** or **not converted** and the threshold is dependent on business.
So, In this task lets go with option 1.

Data pipeline to build user profiles:
Above steps are defined in functions, that automates process of creating user profile on "*registration_id*"

3. Feature Engineering

Propose a business model:



i. Content based prediction model:

Here we will consider columns of video consumption by user. Now this helps to find patterns and conversion predictions. For example: total duration spent on each video, completed 25% [yes or no]?, completed 75% [yes or no]?. Clustering may be required and different algorithms can be performed depending on business.

Final goal of this model is to find if there are any patterns existed in converted Vs not-converted user. Neural Networks performs best in finding patterns enabling signals in neurons in each layer.

ii. Time sensitive prediction model:

Here Let's consider time stamp in dataset and create a new column: "Trial_day" in 7 day trial period ranging [1-7]

This helps us to find at when conversion is done. For example: is conversion done on 4^{th} day? 5^{th} day? Or end of trial?

Considering columns such as number of videos user started to watch. For example: Count of "video_id" on day 1 and so on.

Alternative for feature selection: **Recursive feature elimination** (RFE) can be used to select top features and rank them accordingly.

Answering to business question:
This combination of two models classify's if user account on a trial period is going to convert or not and on what day the event occurred, range of [1-7]?

4. Model Development

Split data: Training, validation and test set. [% depending on size of data set]

Models:

1. Logistics regression
Algorithm to classify if a user/customer will convert or not. Import from scikit learn library.

2. Random forest
Random forest shows good accuracy even if a large proportion of data is missing. Random forest classifier is used to analyze behavior of our user/customers and classify will convert or not.

3. Gradient boost classifier
This is same as RF, here we use intermediate trees to adjust weights of next trees.

4. Naïve Bayes
The output of this classifier is a probability value between 0 to 1.

5. K-Means clustering (k=2)
This uses Euclidean distance to form two clusters. We can visualize how these two groups are varying in distance on a graph.

6. RNN
LSTM and GRU are two **deep learning algorithms** that we can perform to train data and find hidden patterns using **TensorFlow**. GRU is less complex and faster than LSTM. LSTM performs best with large data sets. Transfer learning is other approach that we could use by freezing layers and adding top layers

5. Model Evaluation

Using MLflow:
Now, import mlflow and train different models. This opens in browser as UI that shows all metrics (accuracy scores) and parameters. Keeps track of model versioning's. select model that performs better. Finally register model and process model from staging to production using model registry in MLflow.

Improvements:
Implement RFM (Recency, Frequency and Monetary) for customer segmentation.

Ref Table:

Column Name	Description	Example
registration_id	Unique Registration ID	111111
trial_start_dt	Trial Start Date	10/12/2021
signup_plan_cd	Sign Up Plan Code e.g. Limited Commercial, Commercial Free	Commercial Free
signup_device_cd	Device used to sign up: OTT, Mobile, Tablet	OTT
signup_coupon_cd	Coupon used by user (default Null)	XXYYZZ
campaign_cd	Campaign ID which drove the user to signup (default: Null)	BlackFridayDeal
subscription_type	Differentiate between new or returning customer i.e. winback	'NEW' | 'WINBACK'
converted_ind	Did the user convert from trial to paid?	True
device_type_nm	Device used while watching content: OTT, Mobile, Tablet	OTT
video_id	Unique Video/Content ID	55555
video_duration_seconds	Total Duration of the Content in seconds	6000
video_25_pct_ind	User Completed 25% of the content?	True
video_75_pct_ind	User Completed 75% of the content?	False
video_content_complete_ind	Did user watch the complete content?	False
video_content_type_cd	Content Type: Live or DVR	DVR
video_category_nm	Content Category: Sports,Primetime, etc	Sports
video_genre_nm	Content Genre Name	Action
video_show_nm	Content Show Name	Clarice
video_season_nbr	Content Season Number	1
video_episode_nbr	Content Episode Number	5
profile_nm	Profile Name	User 1
profile_type_cd	Profile Type: Kids, Adults	Adult
birth_year	Birth Year	1990
zipcode	Zip Code	222222