Data Science Work Prompt

VeloCityX

Objectives

Cleaned and processed velocity user engagement data to identify users more likely to purchase virtual merchandise. Analyzed correlations between user activities during race events and their interactions with merchandise purchases and sponsorships. Applied predictive modeling techniques to uncover key insights into user behavior.

Based on the analysis, proposed a new fan challenge aimed at increasing engagement and monetization, with predicted outcomes for both

Steps Involved

1) Understanding Data
2) Correlation
3) K-mean Clustering
4) Implementing Random Forest
5) Proposed fan base challenge

Step 1) Understanding the Data

The data provided consists of 7 features which are follows

- User ID
- Fan Challenges Completed
- Predictive Accuracy in Challenges
- Virtual Merchandise Purchases
- Sponsorship Interactions (Ad Views, Click-Through Rates)
- Time Spent on "Live 360" Coverage
- Real-Time Chat Activity

The dataset contains **100 records and 7 features**. It has no missing or duplicate values, and no preprocessing is required before implementing K-means clustering and modeling. All features are numeric, except for one categorical feature, which is the 'User ID'.

```
df.shape
```

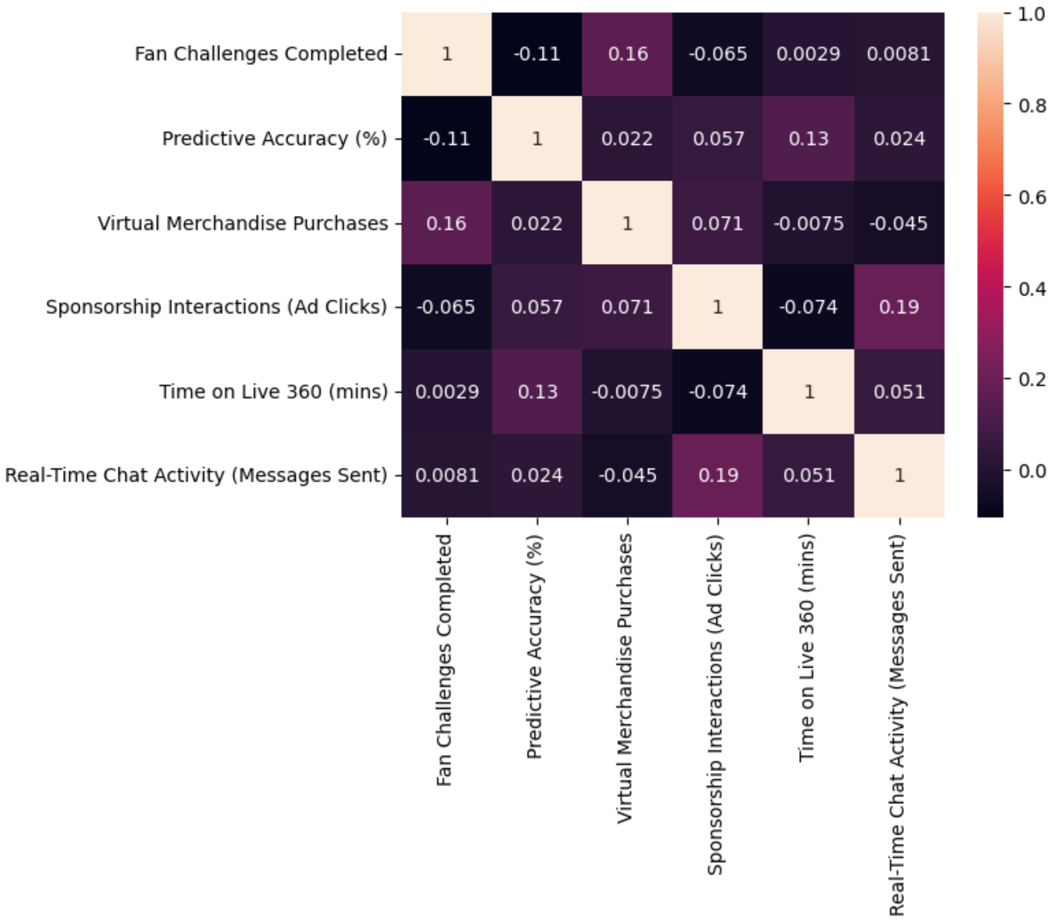
```
(100, 7)
```

```
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 100 entries, 0 to 99
Data columns (total 7 columns):
 #   Column                                  Non-Null Count  Dtype
---  ------                                  --------------  -----
 0   User ID                                 100 non-null    object
 1   Fan Challenges Completed                100 non-null    int64
 2   Predictive Accuracy (%)                 100 non-null    int64
 3   Virtual Merchandise Purchases           100 non-null    int64
 4   Sponsorship Interactions (Ad Clicks)    100 non-null    int64
 5   Time on Live 360 (mins)                 100 non-null    int64
 6   Real-Time Chat Activity (Messages Sent) 100 non-null    int64
dtypes: int64(6), object(1)
memory usage: 5.6+ KB
```

Step 2) Correlation

Creating a Correlation will help understand the relationship of all the columns with each column. This will give an idea which features are related and increases when its correlated features also increase



Key Findings

Fan challenges Completed moderately correlated with **virtual Merchandise Purchases**. That means users who are taking a greater number of fan challenges are more likely to purchase virtual

Sponsorship Interactions are moderately correlated with **virtual Merchandise Purchases**. Which Means user interacting with more

Real-time chat correlated with sponsorship interactions

Step 3) Implementing k-mean clustering

K-means is an unsupervised machine learning algorithm that forms clusters based on similarities among data points. In simple terms, it groups together data points that are similar each other. This algorithm is particularly useful for customer segmentation, grouping people, or identifying patterns within datasets. **For the VelocityX problem statement, applying K-means clustering is an optimal approach for uncovering meaningful user segments**

```
features = df1[['Fan Challenges Completed', 'Predictive Accuracy (%)', 'Sponsorship Interactions (Ad Clicks)',
                 'Time on Live 360 (mins)', 'Real-Time Chat Activity (Messages Sent)']]

# before k-mean implementation standardizing data for better results
s = StandardScaler()
scaled_features = s.fit_transform(features)
```

```
kmeans = KMeans(n_clusters=3, random_state=42)
clusters = kmeans.fit_predict(scaled_features)
# creating a new column for clusters
df1['Cluster'] = clusters
```

Three clusters are formed which are as follows

Cluster 1 - Users who are highly engaged in fan challenges and sponsorship interactions but have low accuracy.

Cluster 2- Users spending more time on live 360 and are active in real chat.

Cluster 3 - Users purchasing virtual merchandise and have high accuracy.

Cluster distribution

```
pca = PCA(n_components=2)
pca_features = pca.fit_transform(scaled_features)


plt.figure(figsize=(10, 7))
plt.scatter(pca_features[:, 0], pca_features[:, 1], c=df1['Cluster'], cmap='viridis', s=50)
plt.title('K-Means Clustering with 2D PCA')
plt.colorbar(label='Cluster')
plt.show()
```

Implementing Cluster Profiling

Cluster profiling will profile detailed **insight** about the cluster like avg value, count of users in each clusters and much more.

Below is the code and its result

```
cluster_profile = df1.groupby('Cluster').mean()

cluster_profile['Cluster Size'] = df['Cluster'].value_counts().sort_index()
print(cluster_profile)
```

```
         Fan Challenges Completed  Predictive Accuracy (%)  \
Cluster
0                        4.421053                70.894737
1                        6.378378                81.243243
2                        5.886364                71.500000

         Virtual Merchandise Purchases  Sponsorship Interactions (Ad Clicks)  \
Cluster
0                             2.157895                             15.684211
1                             2.810811                              5.027027
2                             2.772727                              8.727273

         Time on Live 360 (mins)  Real-Time Chat Activity (Messages Sent)  \
Cluster
0                     153.052632                                38.368421
1                     158.513514                                20.081081
2                      94.590909                                23.477273

         Cluster Size
Cluster
0                  19
1                  37
2                  44
```

Clusters 0, 1, and 2 contain **19, 37, and 44** users, respectively. It is evident that users in Cluster 2 show significantly higher interaction across all metrics. Therefore, the company should prioritize this group to increase merchandise sales. The accompanying Excel tables further illustrate the cluster distribution and provide a clearer understanding of user behavior.

Row Labels	Sum of Virtual Merchandis	Count of User ID	Average Time on Live 360 (mins)	Sum of Fan Challenges Completed
0	41	19	153.0526316	84
1	104	37	158.5135135	236
2	122	44	94.59090909	259
Grand Total	**267**	**100**	**129.35**	**579**

Users from Cluster 0 have contributed only 41 merchandise purchases, while users from Cluster 1 and Cluster 2 have contributed 104 and 122, respectively. This indicates that users in Clusters 1 and 2 have a higher likelihood of purchasing merchandise, suggesting that these segments should be prioritized for marketing efforts

Column3	Column4
Cluster count	Sales
19	15%
37	39%
44	46%
100	100%

This is the contribution of each cluster in the sales of merchandise this makes it more clear that cluster 1 and 2 are more important for the company

Cluster	Real Time Chat Activit	Column2	Column3
Cluster	Min	Max	Avg
0	22	49	25.09
1	0	49	23.19
2	0	48	25.05

Clusters	Live 360	Column1	Column2
Clusters	Min	Max	Avg
0	108	199	130.65
1	106	196	131.3
2	120	120	120

Clusters	FanBase chalenge	column1	column2
	Min	Max	Avg
0	1	9	5.84
1	1	10	5.86
2	5	5.4	5.4

"The table above highlights the top three metrics that influence a user's likelihood of purchasing virtual merchandise. It also shows the minimum, maximum, and average values for each cluster. Users who spend more time watching live 360 streams and interacting with the chatbot are more likely to buy merchandise.

However, it's important to note that these results may be misleading or appear too promising due to the small dataset size, which can lead to inaccurate conclusions.

Steps 4) Implementing predictive modelling

A Random Forest Classifier is used to predict whether a user will purchase merchandise. This method is effective for both classification and regression tasks. It constructs multiple decision trees and combines their outputs to enhance performance. The Random Forest Classifier excels in segmentation and predicting customer purchasing behavior. Therefore, **in the case of the VelocityX app, utilizing a Random Forest Classifier is the optimal choice.**

Benefits of using Random Forest classifier

- RFC does not overfit or underfit the model
- RFC can handle missing data well and maintain accuracy

Splitting the data into training and testing datasets is the best practice for enhancing model performance. The data set consists of users who have purchased merchandise and those who have not, **with 80% of users being buyers and 20% being non-buyers**. It is crucial to ensure that both **buyer categories** are represented equally, or nearly equally, in both the training and testing sets. This balance allows the model to learn effectively and produce accurate results. To achieve this stratified splitting, the **stratify parameter** is passed during the train_test_split initialization.

```python
# Splitting the data into training and testing for more better performance of the model
X = df[['Fan Challenges Completed', 'Predictive Accuracy (%)', 'Sponsorship Interactions (Ad Clicks)',
          'Time on Live 360 (mins)', 'Real-Time Chat Activity (Messages Sent)']]
y = df['Virtual Merchandise Purchases'] > 0


X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.3, random_state=42, stratify=y)

rf = RandomForestClassifier(random_state=42)
rf.fit(X_train, y_train)

y_pred = rf.predict(X_test)
```

Virtual Merchandise Purchases	proportion
True	0.8
False	0.2

dtype: float64

```python
y_train.value_counts(normalize=True)
```

Virtual Merchandise Purchases	proportion
True	0.8
False	0.2

dtype: float64

Model performance

The model's performance was on point as the data is very small model is performing exceptionally well. Let's see the Model performance metrics

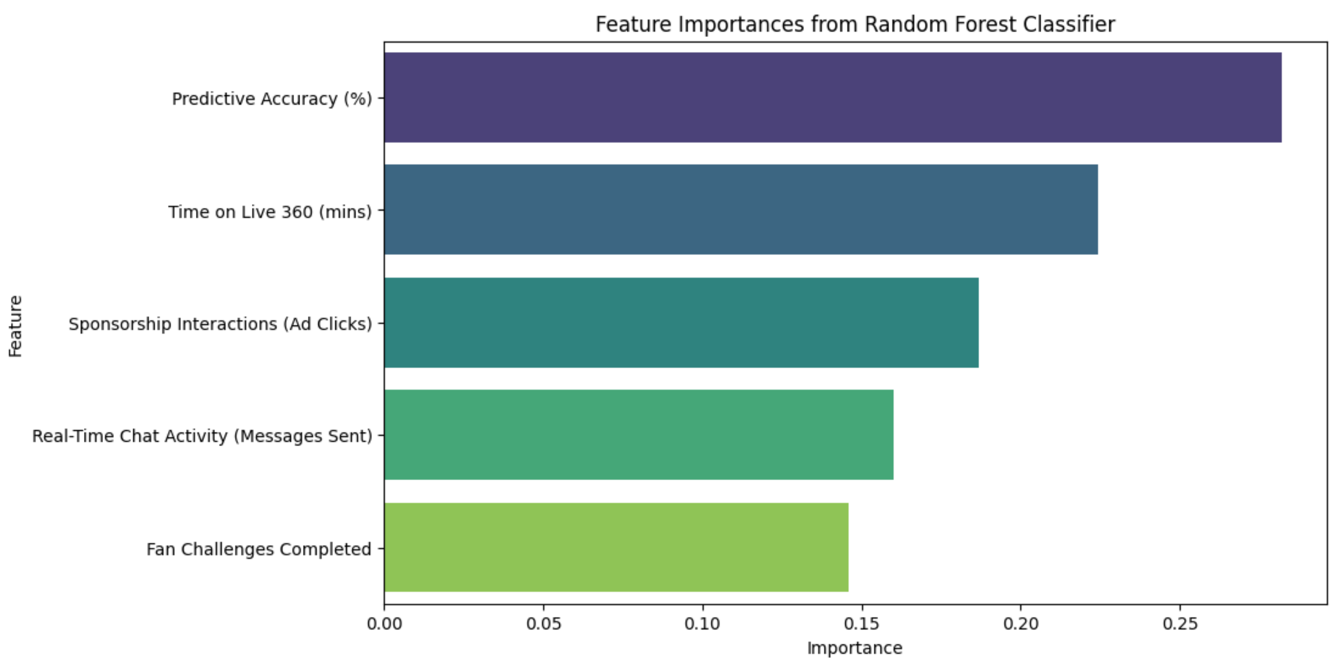
```
[ ]  print(classification_report(y_test, y_pred))
```

```
              precision    recall  f1-score   support

       False       0.00      0.00      0.00         6
        True       0.78      0.88      0.82        24

    accuracy                           0.70        30
   macro avg       0.39      0.44      0.41        30
weighted avg       0.62      0.70      0.66        30
```

It is important to understand the **importance of all the features with respect to virtual merchandise** which is the target variable



Key insights

It is now clear that users who **engage in fan base challenges and live 360 experiences** are more likely to **purchase merchandise**. Additionally, users who interact with the **chatbot and click on sponsor ads** also show a higher likelihood of making purchases.

Interesting insights emerged when analyzing the **important features** related to the target variable. It was found that the most significant feature influencing merchandise purchases is **the prediction accuracy associated with fan challenge participation**. The top three important features identified are **prediction accuracy, live 360 engagement, and ad clicks.**

Therefore, it can be concluded that the company should focus on **increasing fan challenges** to encourage more participation, thereby enhancing prediction accuracy. Users who spend more **time watching live 360 streams, clicking on ads, and participating in fan challenges are more likely to buy merchandise.**

The results may be misleading due to the small size of the dataset. It is likely that the outcomes would differ with a larger dataset. Therefore, using this method with such limited data may not yield the most reliable insights

Step 5) Proposed fan base challenge

Based on our analysis, users who spend more time watching Live 360 are more likely to purchase virtual merchandise. To capitalize on this insight, VeloCityX can introduce a new interactive game called **Live 360 Fan Quest**.

Challenge Description:

Live 360 Fan Quest is designed to engage users by **integrating challenges and real-time quizzes during Live 360 coverage**. This will enhance the viewing experience and significantly boost user engagement

According to my analysis, users spending **more time watching live 360 end up buying virtual merchandise**. So virtualX can introduce a new game called live 360 fan quest where user can play challenges or real time quiz while watching live 360 this can increase user engagement in a large manner. The company can run ads for the same which will increase more users. By this the company has a chance to increase their sales of merchandise. This can be a targeted engagement and can help the company in the long run

Key Features:

Real-Time Quizzes:

Users answer quiz questions related to the race, drivers, and teams during Live 360 coverage.

Points are awarded for correct answers, and a real-time leaderboard keeps users competitive.

Interactive Polls:

Users participate in polls about ongoing race scenarios, such as predicting the next pit stop or driver dual outcomes.Instant poll results foster a sense of community and shared excitement

This challenge will increase **user engagement** as according to my analysis, Live 360 is the most important feature. The challenge will also help in **increasing the sales** of virtual merchandise.