Dating Application (LOVOO) ML Project

My project is based on the work of Jeffrey Mvutu Mabilama, who decided to gather data in dating app called *Lovoo* to try to answer questions in relation to social metrics. What makes a successful profile? Who will have more matches? What are attractive qualities of a profile?

Using Machine Learning models we can learn what constitutes a popular profile and try to predict how many views a profile will receive based on the available data of a user. The views is represented by column ‘profileVisits’.

Disclaimers about the data:

Mabilama collected 2015 data, while the *Lovoo* was expanding in European countries. The metrics of the *Lovoo* include social media like features such as ‘count kisses’, ‘count fans’, that mirror likes and followers. Also, the ‘profileVisits’ column is usually difficult to obtain.

The data gathered by browsing through profiles and searches **recommended** by *Lovoo*'s algorithm for 2 profiles created for this purpose a male, open to friends & chats & dates. That is why there are only female users in the dataset. Also, the recommendation algorithm’s output produced the same set of user profiles which could mean that it was relying on location (to recommend more people nearby than people in different places or countries). This lowered the number of different user profiles that would be presented and included in the dataset.

Data Preparation:

There were two csv files: users\_api\_results and user instances. Upon exploring, there are a number of columns that are duplicates and have different names, mostly boolean values:

* location= locationCity
* isFlirtstar=flirtstar
* isHighlighted = highlighted
* isMobile= mobile
* isOnline= online
* birthd= hasBirthday
* LastOnlineDate= LastOnline
* LastOnlineTime = lastOnlineTs
* isNew = Freshman
* isVip=isVIP
* IsVerified = verified

There were other mutual columns ('gender','age','name','counts\_details','counts\_profileVisits', 'distance', 'flirtInterests\_chat', 'flirtInterests\_date', 'flirtInterests\_friends', 'isSystemProfile', 'lang\_count', 'lang\_de', 'lang\_en', 'lang\_es', 'lang\_fr', 'lang\_it', 'lang\_pt','pictureId', 'whazzup', 'locationCity'', 'crypt', 'counts\_kisses', 'counts\_pictures', 'city') and also some unique columns for each csv.

The Following actions were taken:

1. For the most optimal merge of the two csvs all the columns were examined and manipulated:
   1. fillna(0)/ fillna(‘NA’).
   2. Replacing in all columns were there were: True and False to values 0 and 1.
   3. Uniforming the date
   4. Rounding the float columns
   5. Removal of punctuation
2. Merging of the two csvs into one dataset – outer join
3. Dropping duplicates by userId. userId will be the index
4. Converting objects to string
5. Deleting irrelevant columns

EDA- Explanatory Data Analysis:

* Print types of values, shape and info
* Using describe() for statistics
* Creating EDA report
* Exploring all column values and types
* Checking histograms to find whether any of the columns are normalized. Non were.

Cleansing – Managing outliers and nulls

* Scatter plots and boxplots to identify outliers
* Identifying columns that have outliers that change the distribution and don’t change the correlation with the target value ‘counts\_profileVisits’
* Two columns met the criteria for dropping: ‘counts\_fans’ and ‘distance’. ‘counts\_fans’ was not manipulated, because there could be a person who can have substantially more fans than others. Outliers of ‘distance’ had to be dropped.
* KNNImputer method was used for imputation of the missing ‘distance’ values.
* Missing values: deleting columns that have more than 60% missing. The column ‘locked’
* Cleaning columns with adding none or 0 to complete all columns for the next step of model selection.
* The results: 4008 records

Feature Engineering and Feature Selection:

* ‘userId’ is set as the index
* Splitting the date to into month and day, since the year is the same in all value
* Creating word cloud of the column ‘whazzup’
* Label Encoding
* Feature selection with Lasso and Random Forest

**1. Linear Regression**

* **MSE**: 8.68×1068.68 \times 10^68.68×106
* **RMSE**: 2945.95
* **MAE**: 1481.77
* **RMSLE**: 1.7628

**Interpretation**: The Linear Regression model has moderate errors. Its RMSE is 2945.95, meaning the predictions are off by around 2945.95 units on average. MAE is slightly lower, indicating some stability but still significant error.

**2. Decision Tree**

* **MSE**: 8.65×10−38.65 \times 10^{-3}8.65×10−3
* **RMSE**: 0.093
* **MAE**: 0.0028
* **RMSLE**: 0.000113

**Interpretation**: The Decision Tree model has extremely low error values across all metrics, indicating that it might have overfitted the data (it performs exceptionally well on training data but may not generalize well to unseen data).

**3. Random Forest**

* **MSE**: 1.45×1061.45 \times 10^61.45×106
* **RMSE**: 1204.13
* **MAE**: 449.33
* **RMSLE**: 0.4562

**Interpretation**: The Random Forest model performs well with relatively low errors. The RMSE of 1204.13 suggests that its predictions are fairly accurate, and the lower MAE indicates it handles smaller errors better. RMSLE is also low, indicating good performance on a logarithmic scale.

**4. ADABoost**

* **MSE**: 1.45×1061.45 \times 10^61.45×106
* **RMSE**: 1204.13
* **MAE**: 449.33
* **RMSLE**: 0.4562

**Interpretation**: ADABoost has identical performance to Random Forest, suggesting similar prediction accuracy. It might be useful if Random Forest is computationally expensive or if you prefer ensemble methods.

**5. GBM (Gradient Boosting Machine)**

* **MSE**: 3.37×1063.37 \times 10^63.37×106
* **RMSE**: 1834.94
* **MAE**: 974.79
* **RMSLE**: 1.1004

**Interpretation**: GBM has higher errors compared to Random Forest and ADABoost but is better than Linear Regression. It balances between capturing complex patterns and avoiding overfitting, though it has room for improvement.

**6. SVM (Support Vector Machine)**

* **MSE**: 5.24×1075.24 \times 10^75.24×107
* **RMSE**: 7242.20
* **MAE**: 3229.50
* **RMSLE**: 1.8951

**Interpretation**: The SVM model has the highest errors, indicating it did not perform well on this task. The large RMSE and MAE suggest significant prediction errors, and RMSLE is also high, indicating poor performance with large errors.

**Overall Comparison:**

* **Best Performing Models**:
  + **Decision Tree** (though potentially overfitting)
  + **Random Forest** and **ADABoost** (good balance of accuracy and generalization)
* **Worst Performing Model**: **SVM**, with very high errors, suggesting it is not suitable for this particular task.
* **Linear Regression and GBM** are middle-ground performers, with GBM generally outperforming Linear Regression.

In summary, **Random Forest** or **ADABoost** might be your best bet for a model that balances accuracy and generalization, while **Decision Tree** might work exceptionally well if overfitting isn't a concern.