Class Assignment 2
AI-based technology project using ML with Python

Module Title:	Data Analytics and Machine Learning
Module code:	B9FT114
Lecturer:	Ms Nitya Govindaraju
Submitted by:	Gauri Shingane – 20018204
Submission Date:	25th April 2024

TABLE OF CONTENTS

BUSINESS UNDERSTANDING

When applying for a credit card at any bank various checks need to be done from the financial institution's side. This model could help the banks shortlist the most eligible applicants and determine which applicants have higher chances of defaulting on their payments. The goal is to make the approval process easier for the financial body and minimize the risk of default payment for the financial body.

The key stakeholders here would be the financial body and the applicants. The success criteria would be accurately predicting the customers who are worthy of the credit card and minimizing the issuance of credit cards to more risky applicants more likely to default.

DATA UNDERSTANDING

We have imported the data required for our study from Kaggle. We have downloaded the data to local storage and used the pandas *read_csv* function to import data. The data source contains two datasets having the *ID* as the common key feature, so we have joined the two datasets using the key *ID*. The data frame contains 19 features. For ease of understanding, we have renamed the columns and the new data frame is as shown in figure 1.

```
Data columns (total 19 columns):
 #   Column         Non-Null Count   Dtype
---  ------         --------------   -----
 0   ID             286155 non-null  int64
 1   gender         286154 non-null  object
 2   own_car        286154 non-null  object
 3   own_realty     286154 non-null  object
 4   childcnt       286154 non-null  float64
 5   income         286154 non-null  float64
 6   incometp       286154 non-null  object
 7   edutp          286154 non-null  object
 8   fmstatus       286154 non-null  object
 9   housetp        286154 non-null  object
 10  age            286154 non-null  float64
 11  days_employed  286154 non-null  float64
 12  mbl            286154 non-null  float64
 13  wkphone        286154 non-null  float64
 14  phone          286154 non-null  float64
 15  email          286154 non-null  float64
 16  occtp          198506 non-null  object
 17  famsize        286154 non-null  float64
 18  account_age    36105 non-null   float64
dtypes: float64(10), int64(1), object(8)
```

Figure 1: Overview of the data

We then analyse the feature *STATUS* to determine the risk factor for all applicants which will in turn help create the *target* feature for our model.

Next, we need to handle all *Null* values from the data. There are missing values found only in the *occtp* feature, so we remove all records from data with *Null* values. Once initial data cleaning is done, we are left with 10 columns in the data frame of which 19 are determinant features and 1 is the target outcome for our machine learning models.

DATA PREPARATION

Data encoding

- One-hot encoding: We have used one-hot encoding to nominal columns with no specific order and created binary values for these categorical features.
- Ordinal encoding: For categorical features with a specific order in place for the values we have used ordinal encoding.

Feature scaling

We have applied scaling to the numerical continuous features such as *income* and *days_employed* as normalized values for the *famsize* feature. We have also scaled *age* and *days_employed* features to a yearly scale.

Feature selection

The features we have dropped are ID, mbl, childcnt, incometp and account_age. We have dropped these columns because,

- ID does not help predict the outcome
- mbl does not affect the outcome as everyone has a mobile
- childcnt is highly correlated with famsize
- occtp because it does not affect the outcome
- account_age because we have deduced the target variable from it will lead to overfitting

Data Balancing

We have applied Synthetic Minority Oversampling Technique (SMOTE), to address the imbalance in the data. This method helps balance the data by generating synthetic samples for the minority features in the data.

Data Visualization

(a) Analysing binary features: We analyse the binary features to understand the skewness in the data. We have used pie charts for this purpose.

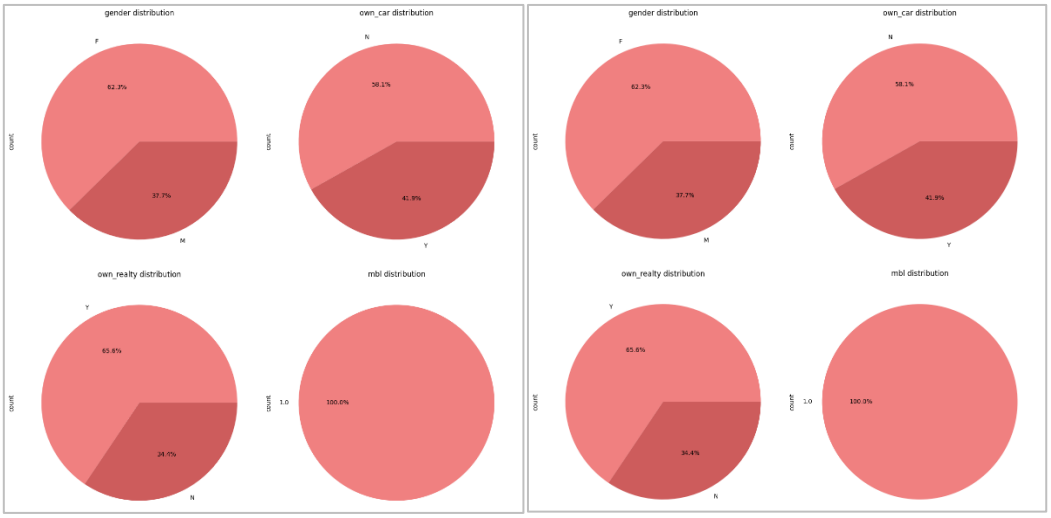

Figure 2: Analysing binary features

From Figure 2 we can see that:

- 'mbl' feature is unfit for analysing the data and may affect the efficiency of our ML algorithms later.
- most of the features are biased

Also, we can observe that most applicants are,

- Females
- own car
- own realty

(b) Analysing categorical features: By analysing categorical features we can understand the relationship between features, identify biases within the data, and analyse the impact of these features on the overall credit card approval decision.

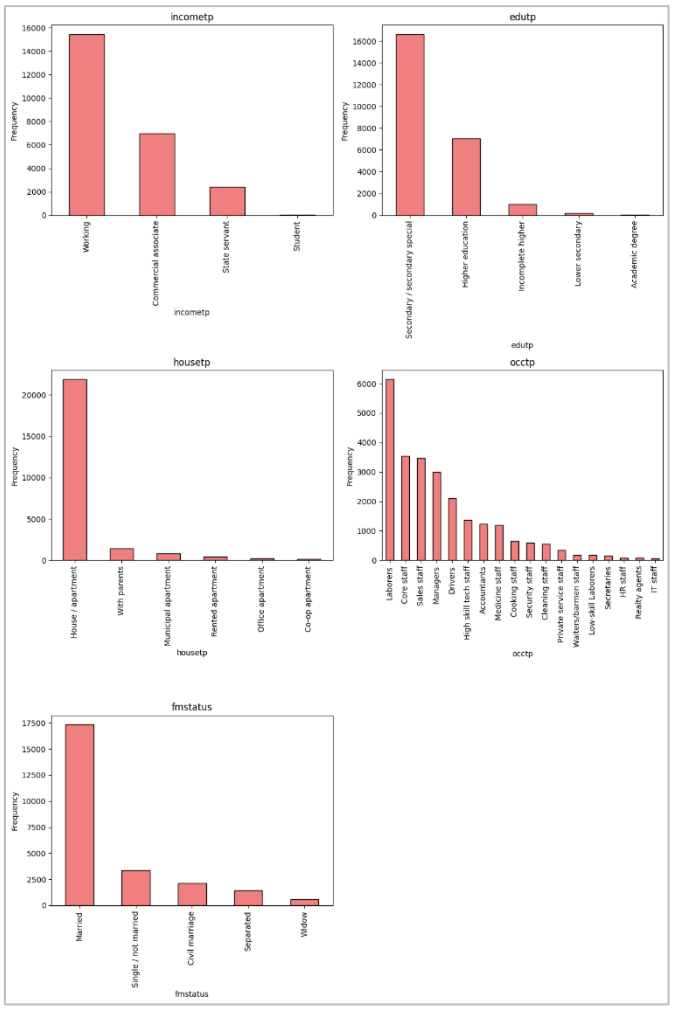

Figure 3: Analysing categorical features

In figure 3, we can see that most applicants are

- married

- working
- and have at least a secondary level of education

(c) Bivariate/Multivariate analysis: We use this method to analyse trends in multiple features concerning each other.

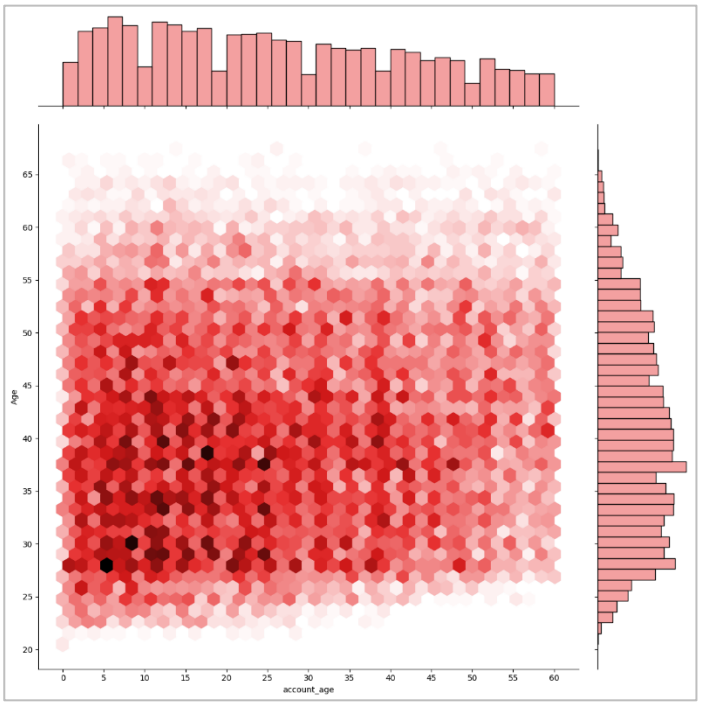

Figure 4: Joint plot for account_age vs age

From Figure, 4 We can observe from the above graph that most credit card applicants are,

- between ages 27 and 45
- have an account since 5 to 25 months

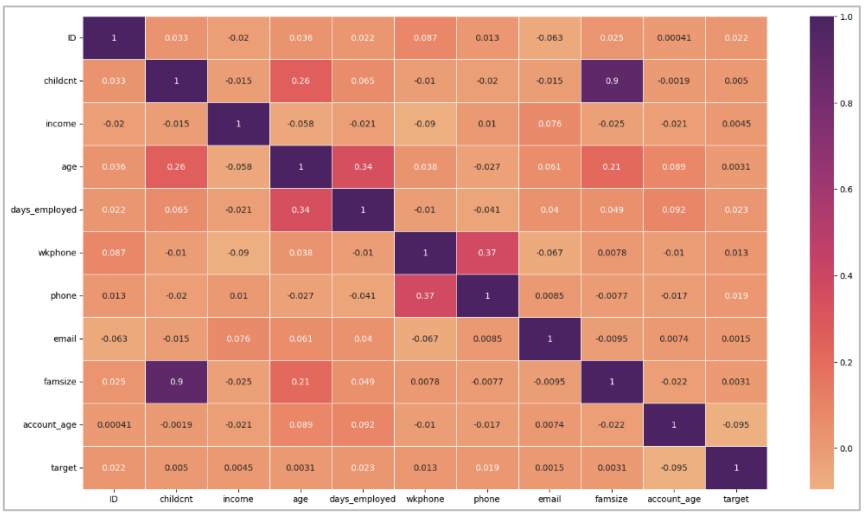

Figure 5: Heatmap for categorical features

From the heatmap in Figure 5, we can observe that,

- no feature is strongly correlated with the target feature
- high correlation between *famsize* and *childcnt*, as more children leads to bigger families
- *age* has a positive correlation with *famsize*, *childcnt* and *days_employed*

From the above visual analysis, we can infer that,

- Typical applicant is a female in her 40's, married, has been employed for about five years, has completed her secondary education, and owns a property.
- Also age and income have minimal or no effect on the target variable
- Most applicants are between 27 and 45 years of age and have accounts from 5 to 25 months.

MACHINE LEARNING MODELS IMPLEMENTATION

Models

Support Vector Classifier finds a border between classes and finds the best fit for the border. The border is called a hyperplane, which ensures the classes are as far away from each other as possible. Support Vector Machines can be used for both classification and regression problems.

Logistic Regression Algorithm uses a sigmoid function to predict the label for data. The LR is used mostly when we expect a binary outcome from the problem. The LR uses the sigmoid function and the probability of the label is calculated and assigned to the data.

Random Forest Classifier is a collection of decision trees that aggregates the results of the individual trees to predict the outcome. Random forest classifier is known for high accuracy when handling classification problems and highly noisy data.

XGBoost is again an algorithm used both in classification and regression problems. It uses multiple decision trees where each tree takes input from the previous tree thus optimizing and correcting any errors made by the previous tree.

Performance evaluation

(a) Logistic Regression: The model predicts credit card approvals with 61% accuracy, but is not much good at differentiating between the target variable. This might be results of imbalanced data or model limitations.

	precision	recall	f1-score	support
0	0.60	0.61	0.61	4880
1	0.62	0.61	0.61	5005
accuracy			0.61	9885
macro avg	0.61	0.61	0.61	9885
weighted avg	0.61	0.61	0.61	9885

(b) Random forest classifier: This model perfectly predicts all the outcome.

	precision	recall	f1-score	support
0	1.00	0.99	1.00	7463
1	0.99	1.00	1.00	7365
accuracy			1.00	14828
macro avg	1.00	1.00	1.00	14828
weighted avg	1.00	1.00	1.00	14828

(c) XGBoost: XGBoost algorithm predicts the approvals with 98% accuracy. It is highly precise but the recall is bit lower for approvals when compared. This models seems to be very efficient in predicting credit card approvals.

	precision	recall	f1-score	support
0	0.96	0.99	0.98	4894
1	0.99	0.96	0.98	4991
accuracy			0.98	9885
macro avg	0.98	0.98	0.98	9885
weighted avg	0.98	0.98	0.98	9885

(d) Support Vector Classifier: The SVC has accuracy of 67% while predicting the credit card approval. The accuracy is not much good even after refinement we only get 66% recall.

	precision	recall	f1-score	support
0	0.70	0.60	0.65	7338
1	0.66	0.74	0.70	7490
accuracy			0.67	14828
macro avg	0.68	0.67	0.67	14828
weighted avg	0.68	0.67	0.67	14828

The overall performance of the machine learning models used for credit card approval prediction varies greatly. Overall, XGBoost appears to be the best-performing model out of the ones we tested. We can retrain XGBoost with a larger dataset or explore techniques to improve the performance of the other models.

	Model	Accuracy	Specificity	Sensitivity	AUC	F1 Score
0	RandomForestClassifier	0.995684	0.994774	0.996606	0.998691	0.995659
1	XGBClassifier	0.977946	0.994279	0.961931	0.997666	0.977800
2	LogisticRegression	0.610015	0.607992	0.611988	0.655087	0.613766
3	BaggingClassifier_LR	0.611836	0.608811	0.614785	0.655251	0.615954
4	SupportVectorClassifier	0.669881	0.605070	0.733378	0.738417	0.691770
5	BaggingClassifier_SVC	0.669881	0.605070	0.733378	0.695836	0.691770

DEPLOYMENT AND CONCLUSION

Deployment plan

The deployment process is when we move our ML model from an offline environment and integrate it into an existing production environment like a live application. There are many ways to deploy ML models like batch deployment, real-time deployment, streaming deployment, and edge deployment. To deploy the above-created credit card approval prediction model first we would need to use a cloud platform like AWS or GCP. We would also need a monitoring infrastructure to make sure it is working as expected.

Results and Conclusion

In this Credit card approval prediction, we have implemented various machine learning models to predict the outcome. We also found some interesting trends and evaluated the performance of all the models.

As an outcome of our EDA, we found that there is a high preference for approval to candidates who are females, who are married, working and have at least secondary education. In terms of the machine learning models, the XGBoost model was the most effective in predicting outcomes and achieved high accuracy rates.

However, we can still improve the performance by improving data quality and exploring techniques to enhance the other machine learning models like Logistic Regression and Support Vector Classifier used in this study.

SOURCES

1) https://www.kaggle.com/datasets/rikdifos/credit-card-approval-prediction/data
2) https://www.linkedin.com/pulse/top-6-machine-learning-classification-algorithms-mrinmoy-paul-/