Capstone Project-3

Credit Card Default Prediction

Prepared by:

Suhail Shaikh

Contents

- Problem statement
- Understanding the data set
- Data cleaning
- Exploratory data analysis
- Understanding correlation
- Model implementation
- Model evaluation
- Conclusion
- Suggestions
- References

Problem Statement

- To build a model to identify whether the credit card applicant will default or not based on his repayment history and other important factors.

- In order to achieve this, we need to develop a supervised learning model using classification algorithms.

Understanding the Data Set

Data set has 30000 rows and 25 columns The columns in data set have information as mentioned below,

- **ID** : Customer id
- **LIMIT_BAL** : Credit limit of customer including his family
- **SEX** : (1 = male; 2 = female)
- **EDUCATION** : (1 =graduate school; 2 =university; 3 =high school; 4 =others)
- **MARRIAGE** : (1 = married; 2 = single; 3 = others)
- **AGE** : Age of customer in Years
- **PAY_0** : Humidity in living room area, in
- **PAY_2** : Temperature in laundry room area

Cont…

- **PAY_0** : Repayment status in September 2005
- **PAY_2** : Repayment status in August 2005
- **PAY_3** : Repayment status in July 2005
- **PAY_4** : Repayment status in June 2005
- **PAY_5** : Repayment status in May 2005
- **PAY_6** : Repayment status in April 2005
- **BILL_AMT1** : Billing statement in September 2005
- **BILL_AMT2** : Billing statement in August 2005
- **BILL_AMT3** : Billing statement in July 2005
- **BILL_AMT4** : Billing statement in June 2005
- **BILL_AMT5** : Billing statement in May 2005

Cont…

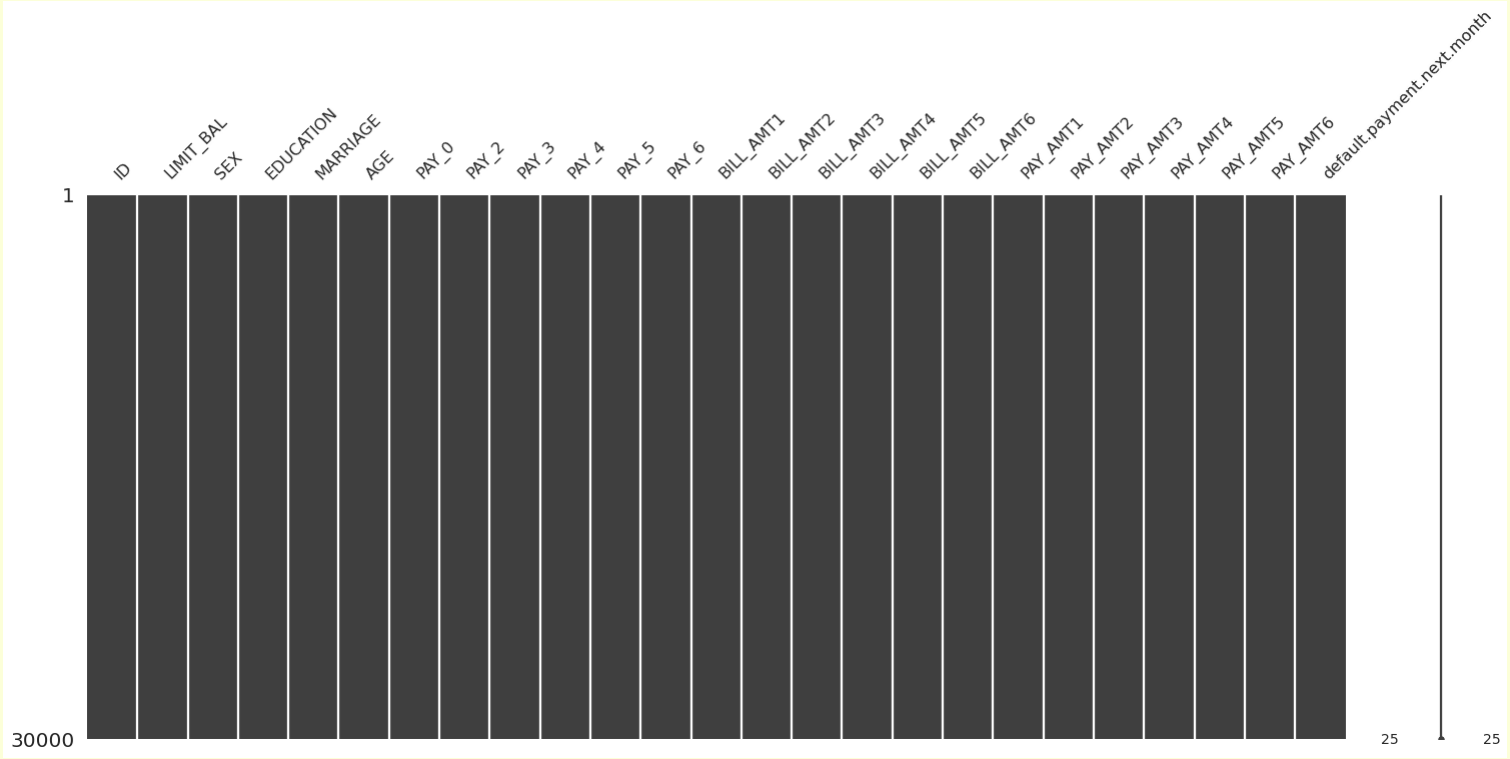
- **BILL_AMT6** : Billing statement in April 2005
- **PAY_AMT1** : Amount paid in September 2005
- **PAY_AMT2** : Amount paid in August 2005
- **PAY_AMT3** : Amount paid in July 2005
- **PAY_AMT4** : Amount paid in June 2005
- **PAY_AMT5** : Amount paid in May 2005
- **PAY_AMT6** : Amount paid in April 2005
- **default payment next month** : default payment (Yes = 1, No = 0)

Data Cleaning

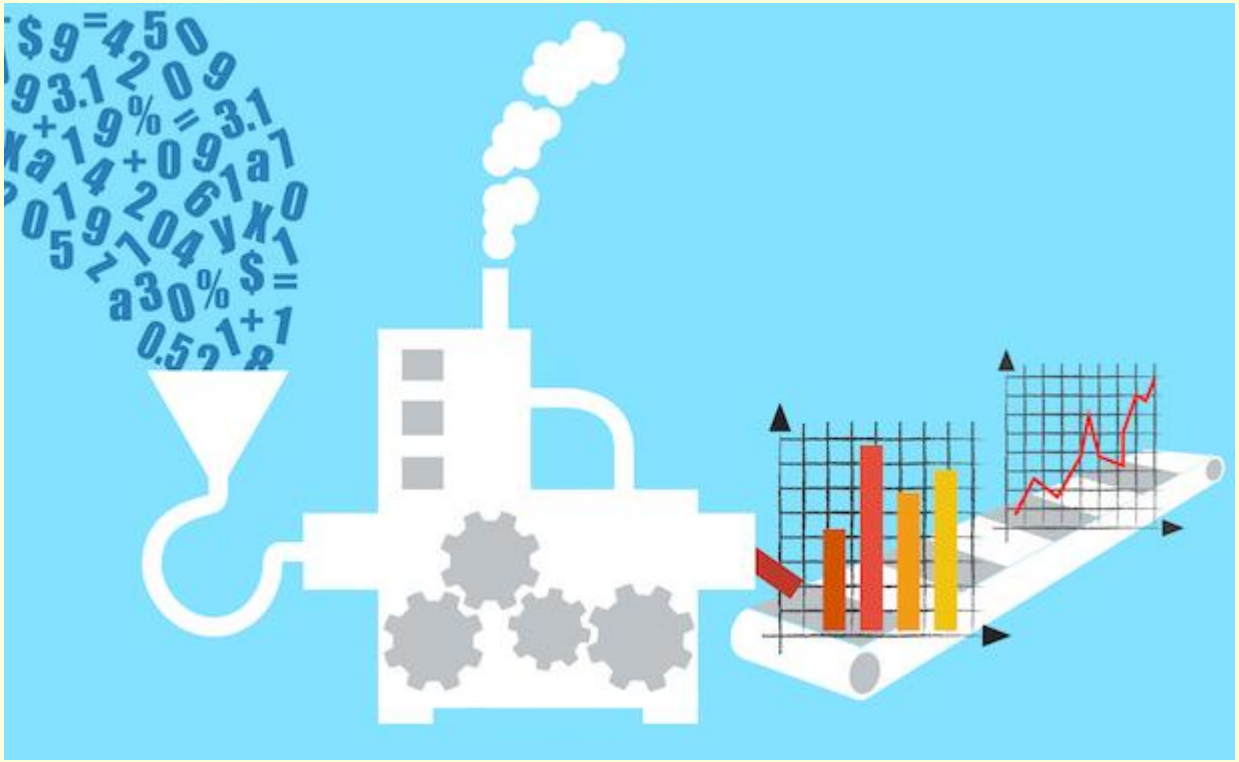
- Data set has no null values

Column	Null Values	Column	Null Values
ID	0	BILL_AMT2	0
LIMIT_BAL	0	BILL_AMT3	0
SEX	0	BILL_AMT4	0
EDUCATION	0	BILL_AMT5	0
MARRIAGE	0	BILL_AMT6	0
AGE	0	PAY_AMT1	0
PAY_0	0	PAY_AMT2	0
PAY_2	0	PAY_AMT3	0
PAY_3	0	PAY_AMT4	0
PAY_4	0	PAY_AMT5	0
PAY_5	0	PAY_AMT6	0
PAY_6	0	default.payment.next.month	0
BILL_AMT1	0	**Total**	0

Visualizing Null Values



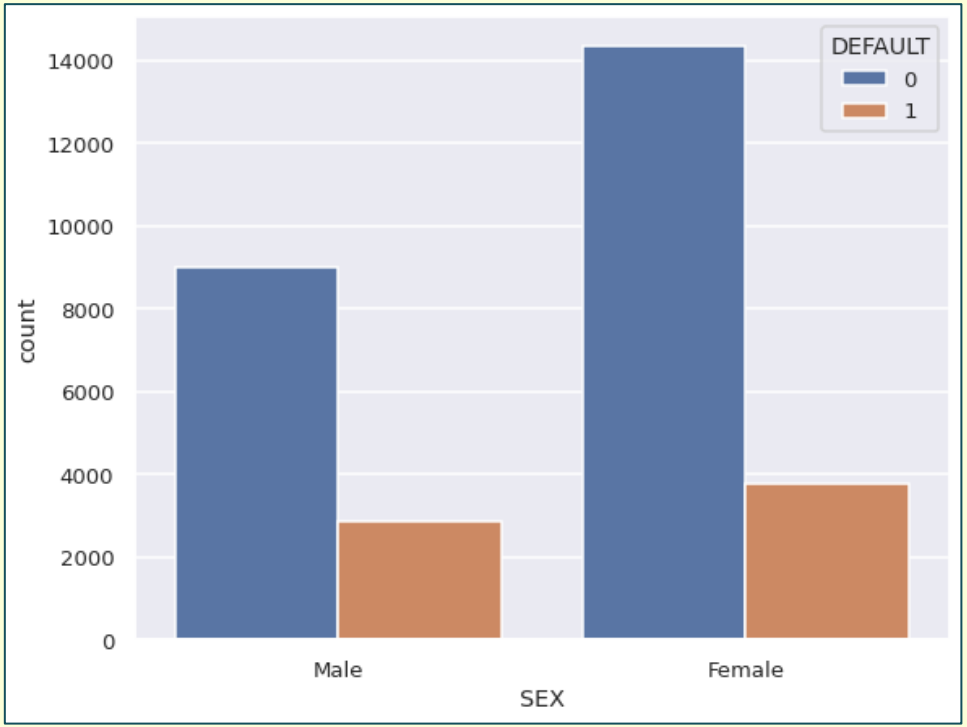
Exploratory Data Analysis



Gender Vs Default

	SEX	NOT_DEFAULT	DEFAULT	TOTAL
0	Male	9015	2873	11888
1	Female	14349	3763	18112

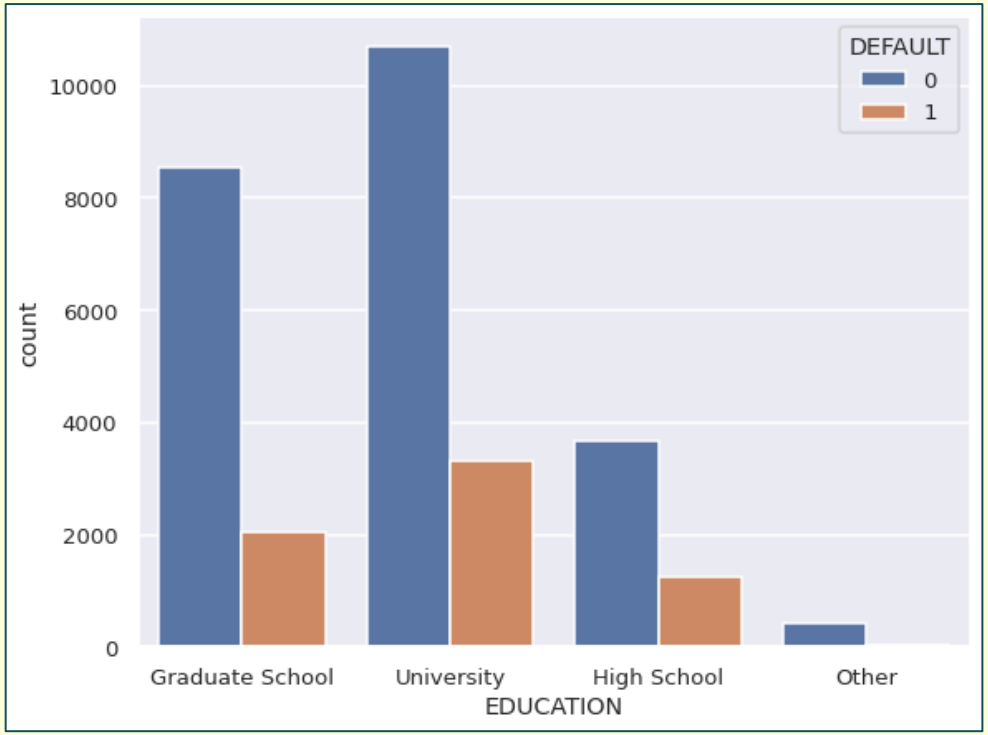
	SEX	NOT_DEFAULT(%)	DEFAULT(%)
0	Male	75.83	24.17
1	Female	79.22	20.78



Education Vs. Default

	EDUCATION	NOT_DEFAULT	DEFAULT	TOTAL
0	Graduate School	8549	2036	10585
1	University	10700	3330	14030
2	High School	3680	1237	4917
3	Other	435	33	468

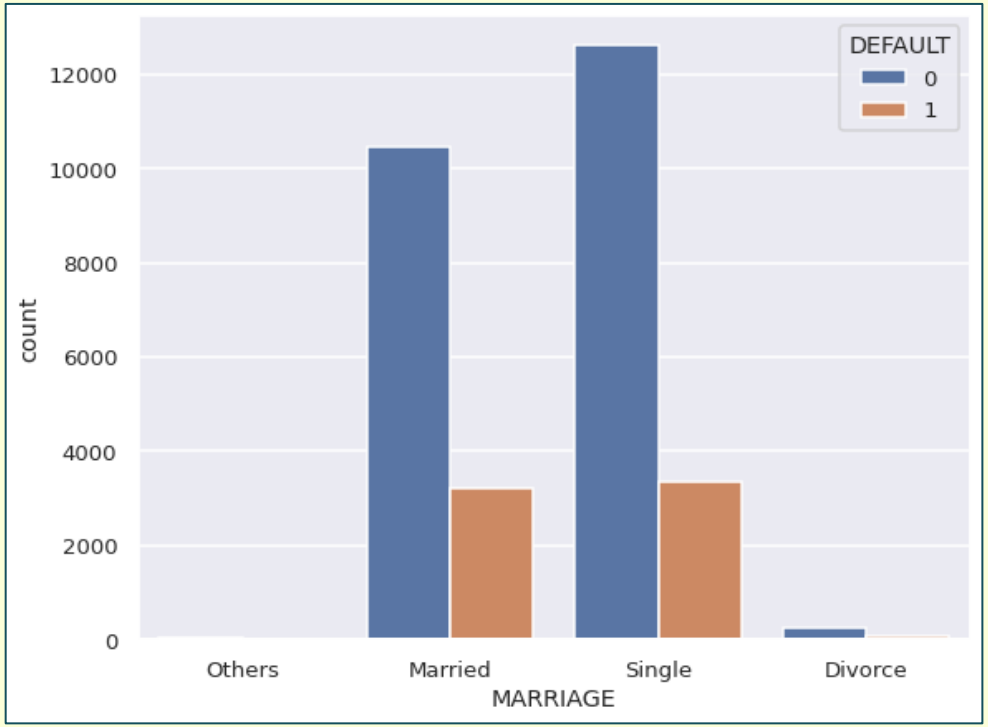
	EDUCATION	NOT_DEFAULT(%)	DEFAULT(%)
0	Graduate School	80.77	19.23
1	University	76.27	23.73
2	High School	74.84	25.16
3	Other	92.95	7.05



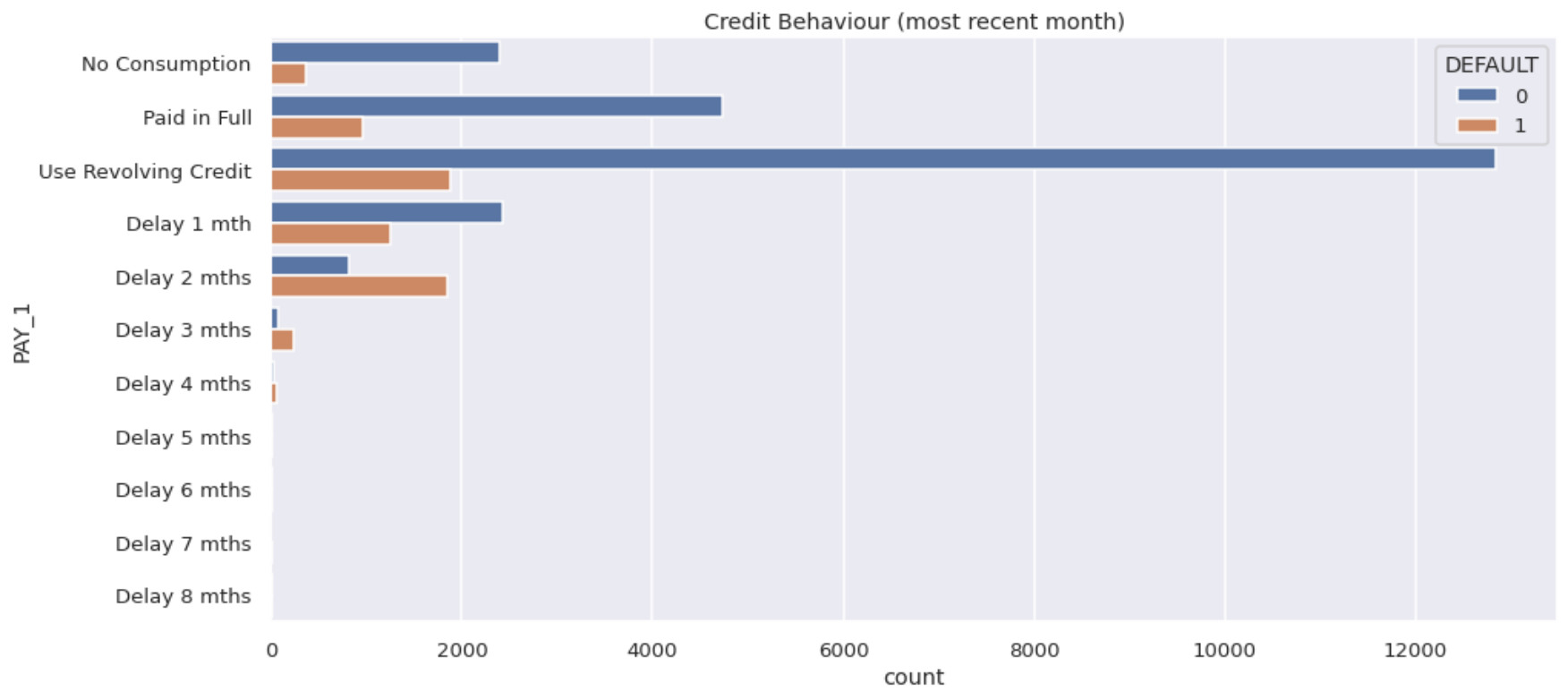
Marital Status Vs Default

	MARRIAGE	NOT_DEFAULT	DEFAULT	TOTAL
0	Others	49	5	54
1	Married	10453	3206	13659
2	Single	12623	3341	15964
3	Divorce	239	84	323

	MARRIAGE	NOT_DEFAULT(%)	DEFAULT(%)
0	Others	90.74	9.26
1	Married	76.53	23.47
2	Single	79.07	20.93
3	Divorce	73.99	26.01



Credit Behavior Vs. Default

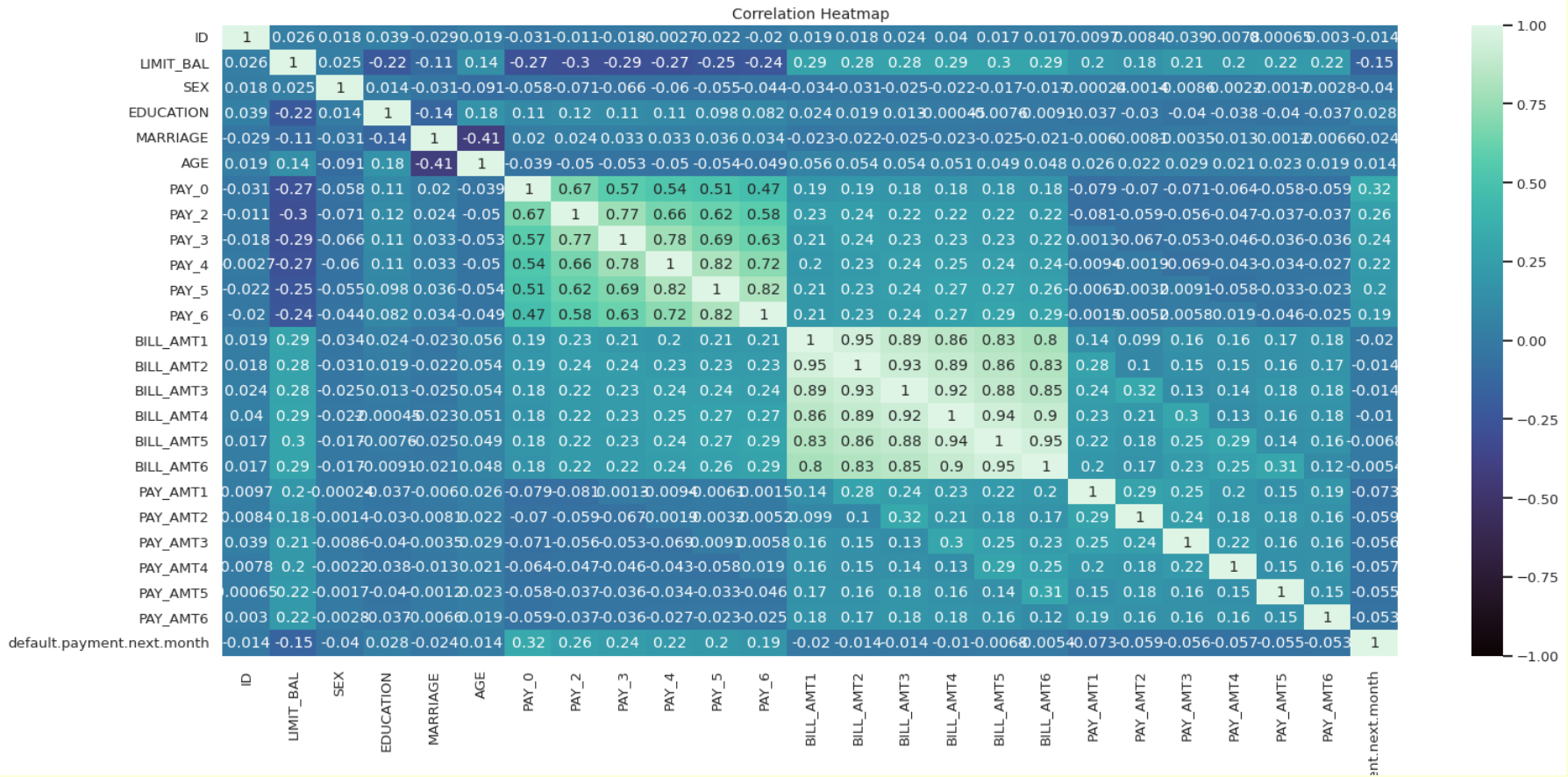


Credit Behavior Vs. Default

	PAY_1	NOT_DEFAULT(%)	DEFAULT(%)
0	No Consumption	86.77	13.23
1	Paid in Full	83.22	16.78
2	Use Revolving Credit	87.19	12.81
3	Delay 1 mth	66.05	33.95
4	Delay 2 mths	30.86	69.14
5	Delay 3 mths	24.22	75.78
6	Delay 4 mths	31.58	68.42
7	Delay 5 mths	50.00	50.00
8	Delay 6 mths	45.45	54.55
9	Delay 7 mths	22.22	77.78
10	Delay 8 mths	42.11	57.89

	PAY_1	NOT_DEFAULT	DEFAULT	TOTAL
0	No Consumption	2394	365	2759
1	Paid in Full	4732	954	5686
2	Use Revolving Credit	12849	1888	14737
3	Delay 1 mth	2436	1252	3688
4	Delay 2 mths	823	1844	2667
5	Delay 3 mths	78	244	322
6	Delay 4 mths	24	52	76
7	Delay 5 mths	13	13	26
8	Delay 6 mths	5	6	11
9	Delay 7 mths	2	7	9
10	Delay 8 mths	8	11	19

Understanding Correlation



Model Implementation

- H

```
[ ]  models = {
         LogisticRegression():        "     Logistic Regression",
         SVC():                       "  Support Vector Machine",
         RandomForestClassifier():    "Random Forest Classifier",
         XGBClassifier():             "     XG Boost Classifier"

     }

     for model in models.keys():
         model.fit(X_train, y_train)
```

- We have used Logistic Regression, SVM Algorithm, Random Forest Classifier and XG Boost Classifier.

Model Evaluation

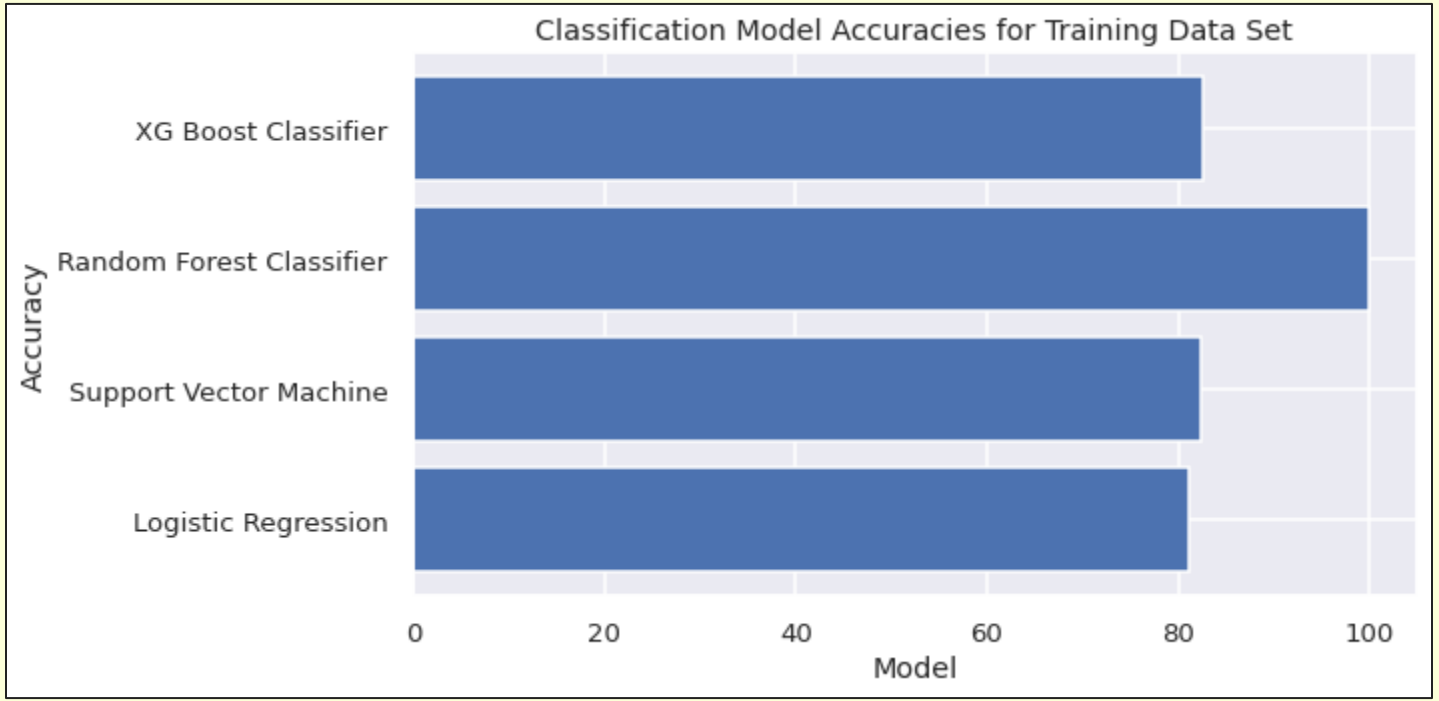
```
Logistic Regression: 81.12%
Support Vector Machine: 82.39%
Random Forest Classifier: 99.93%
XG Boost Classifier: 82.50%
```

For Training Data Set

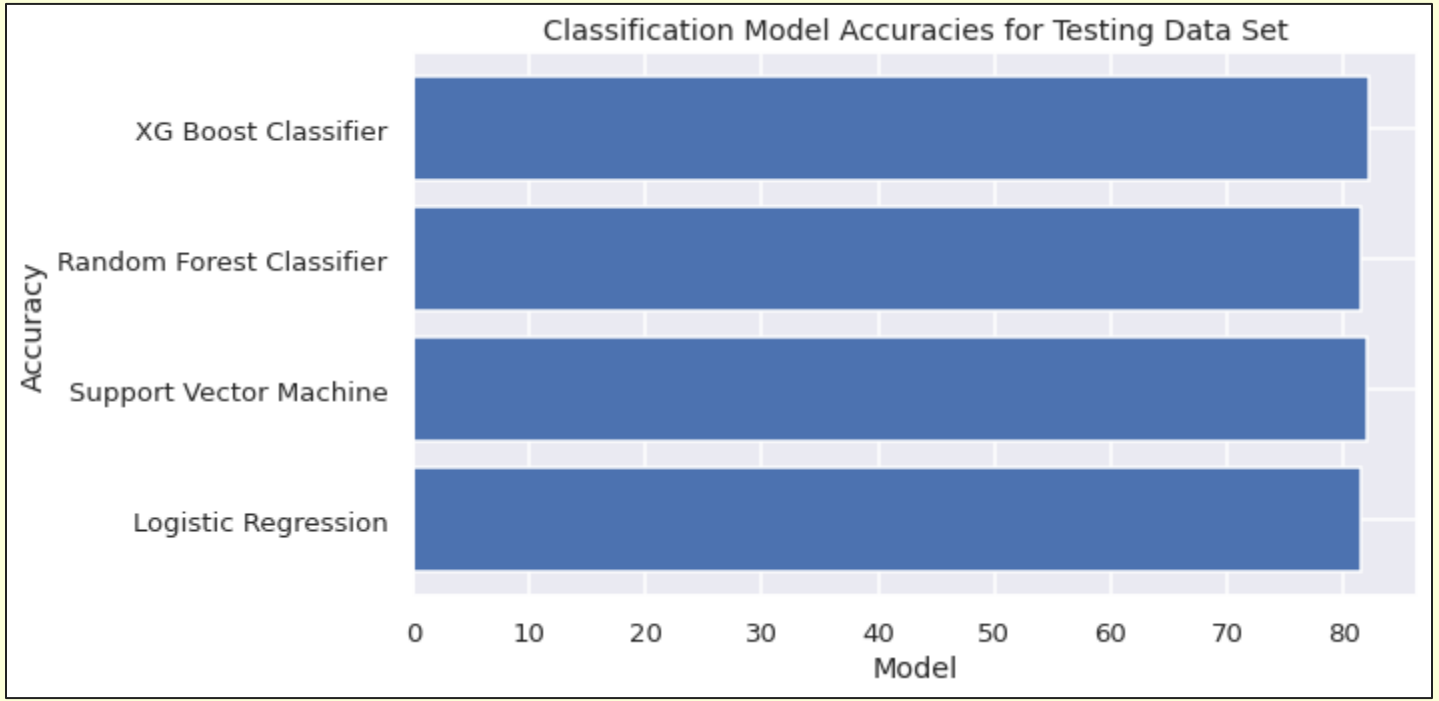
For Testing Data Set

```
Logistic Regression: 81.43%
Support Vector Machine: 82.03%
Random Forest Classifier: 81.56%
XG Boost Classifier: 82.21%
```

- XG Boost Classifier has higher accuracy 82.21% followed by SVM 82.03%



❑ **Model Accuracies for Training Data Set**



❑ **Model Accuracies for Testing Data Set**

Conclusion

- From accuracy results we can conclude that XG Boost Classifier has highest accuracy followed by Random Forest Classifier.
- From exploratory data analysis we can conclude that default rate for educated customer is less.
- Default rate is slightly higher in Male Customers as compare to Female Customers.
- Default rate is higher for married and divorced customers as compared to single customers.
- When payment is delayed more than 2 months, the chances of default goes higher than 50%.

Suggestions to Improve Credit Card Sales in Banks

- To improve transactions, we should offer credit cards to highly educated customer based on their credit score and transaction history.
- We can offer credit cards to customers having no delay in their repayment history.
- Customers between age group 25 to 50 have higher credit card utilization, so to improve transaction we should offer credit cards between 25 to 50 age group.

References

- Kaggle
- Github
- Youtube
- Towards Data Science
- Code Basics
- Analytics Vidya
- Stack Over Flow

Thank You