IE 7275 Data Mining in
Engineering Group36

CREDRISK: A CREDIT RISK PREDICTOR MODEL USING MACHINE LEARNING

Yashasvi Sharma
Pranjal Kalekar

IE 7275 Data Mining in Engineering
Group 36

CREDRISK

Yashasvi Sharma
Pranjal Kalekar

Agenda

- Project selection
- Problem definition
- Project Objective
- Data collection
- Data exploration, visualization, and processing
- Dimension reduction, variable selection
- Model exploration
- Model selection
- Model performance evaluation
- Performance visualization

PROBLEM DEFINITION

- One of the primary challenges for lenders is to assess the creditworthiness of borrowers. The traditional credit score models have limitations in accurately predicting default risk.
- lenders need a reliable and accurate credit risk assessment model to make informed decisions and manage their risk exposure
- Todays lenders need:
  - Better Access to Credit
  - Reduced Risk of Default
  - Lower Interest Rates
  - Better Financial Health

Project Objective

The project aims to develop a reliable and accurate credit risk assessment model using Lending Club data that can help lenders make informed decisions about lending and manage their risk exposure.

Data collection

LENDING CLUB

Expanding financial opportunities for all Americans through responsible innovation

LendingClub is a peer-to-peer lending service provider that allows individual investors to partially fund personal loans as well as buy and sell notes backing the loans on a secondary market.
LendingClub makes historical data available to the public via an API.
Financial datasets containing information on credit history, income, employment status, and other relevant data from a sample of individuals or organizations.

DataDescription

1

The data sets fare of year 2020-2021

2

143 attributes in total
1 target attribute, 'loan status'.

3

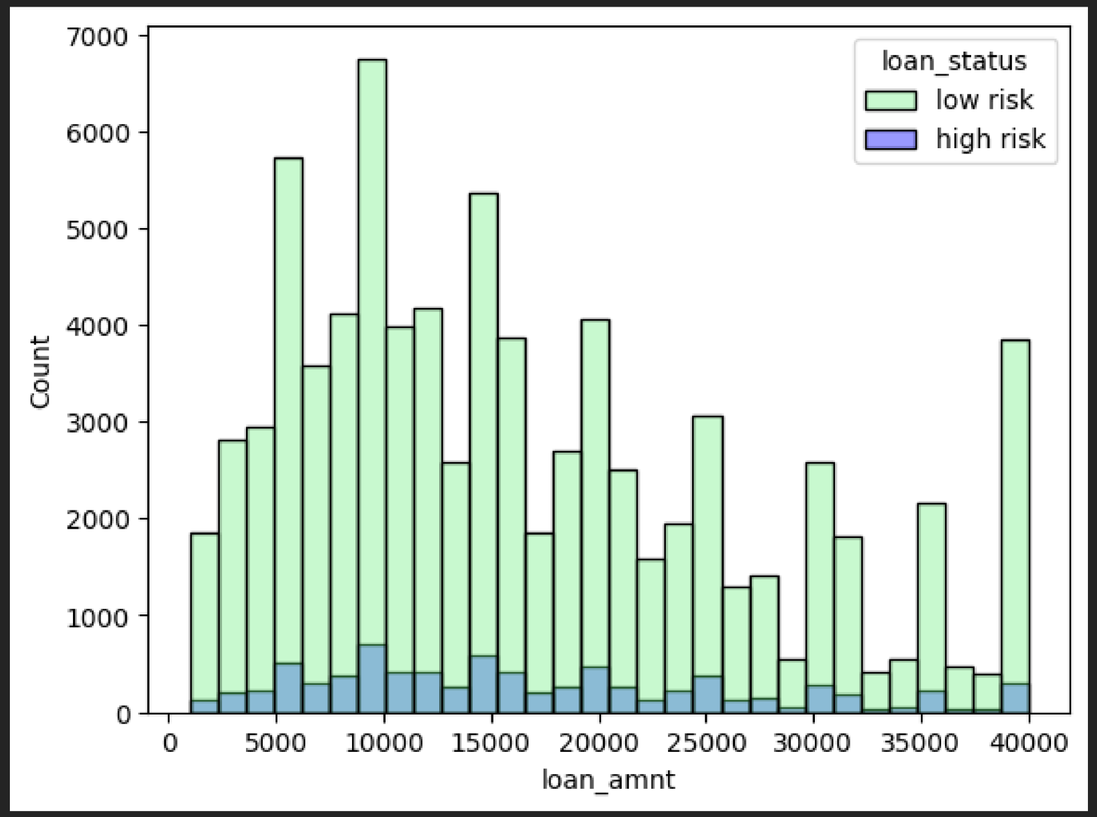
107 numerical attributes and 37 categorical attributes

4

finalized 69 columns after data cleaning and feature engineering

Back to Agenda Page

Data exploration and visualization



HistPlot of loan_amnt and ‘count’

High Risk Loans in each state

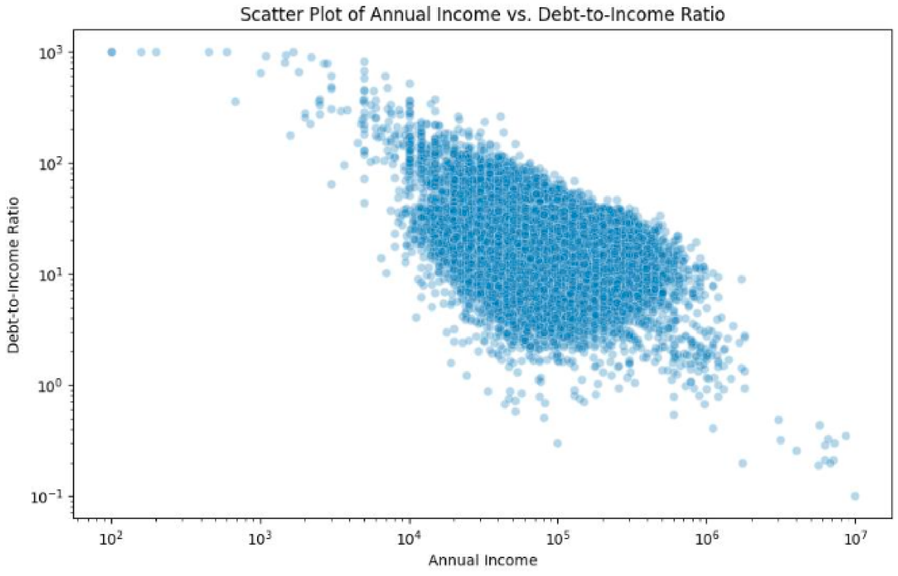


Fig 3.1.5: Scatter Plot of Annual Income vs. Debt-to-Income Ratio

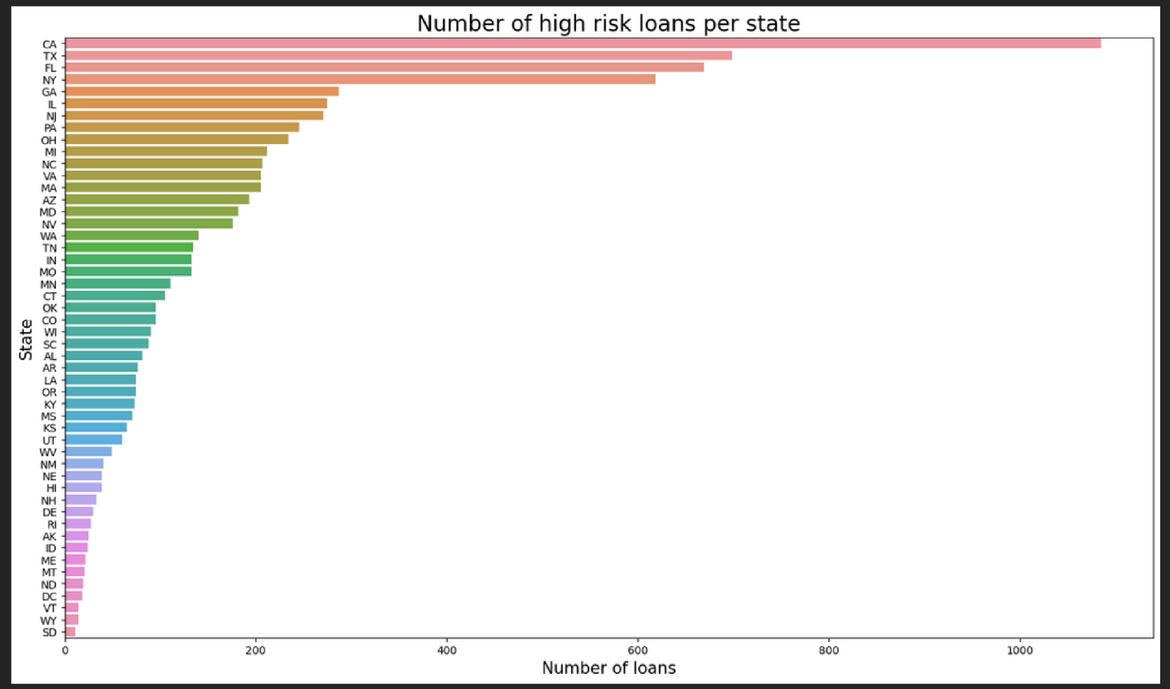
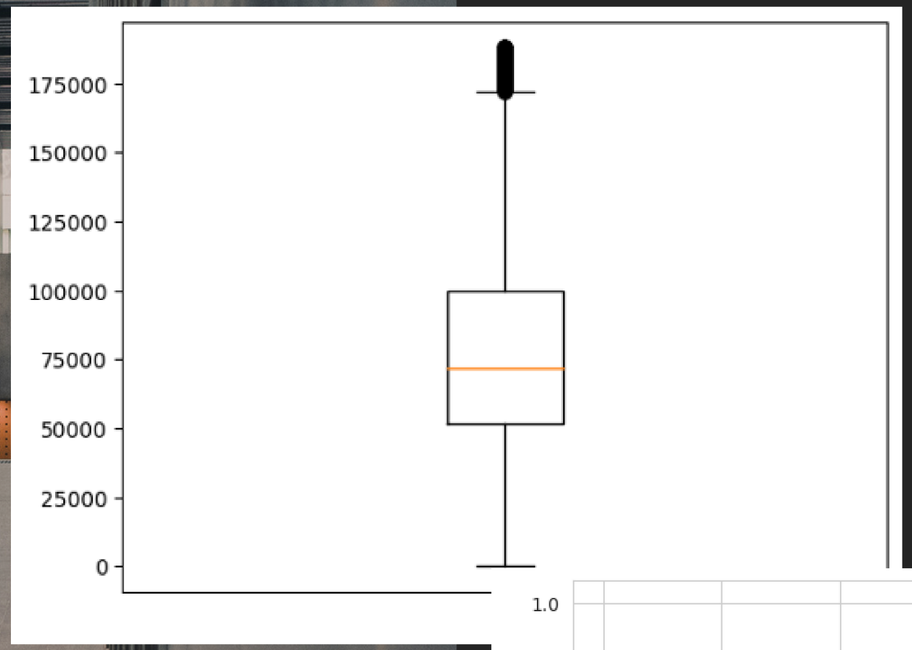
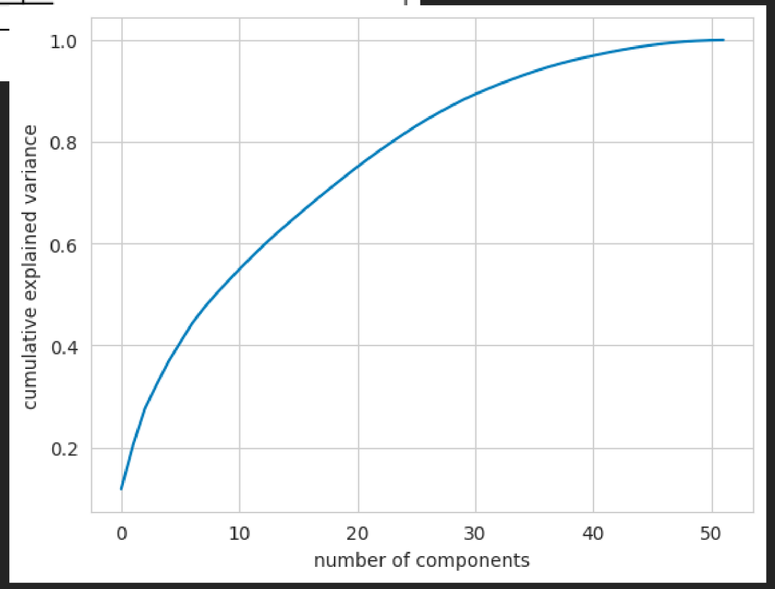


Fig 3.1.3: Bar Plot of high risk loan per state

Data exploration and visualization
Correlation Matrix

DIMENSION REDUCTION AND VARIABLE SELECTION





PHASE 1

Removing Columns with 75 % of missing values and no unique values

PHASE 2

Removing highly correlated Columns

PHASE 3

Filling na values and removing outliers

PHASE 4

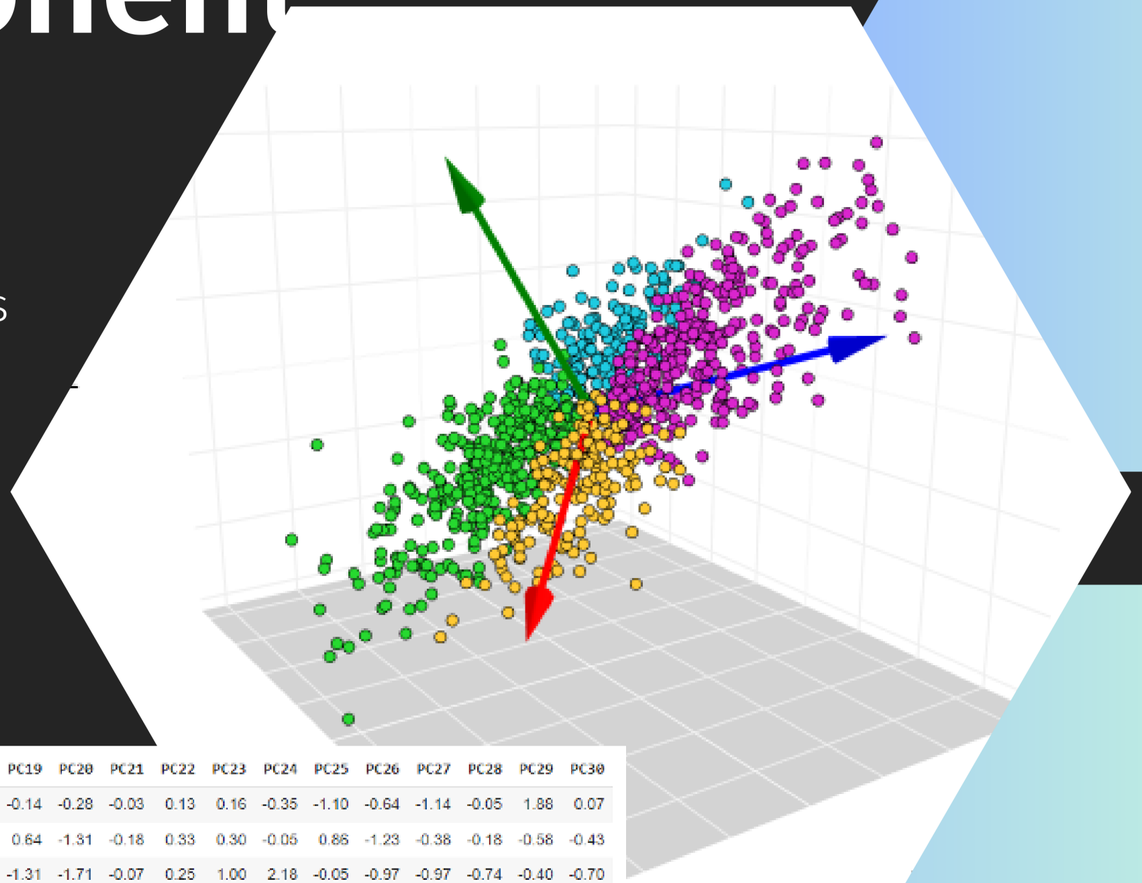
PCA

PHASE 5

Chi-squared

Principal Component Analysis (PCA)

- PCA is to represent a multivariate data table as smaller set of variables (summary indices) in order to observe trends, jumps, clusters and outliers.
- This overview may uncover the relationships between observations and variables, and among the variables.

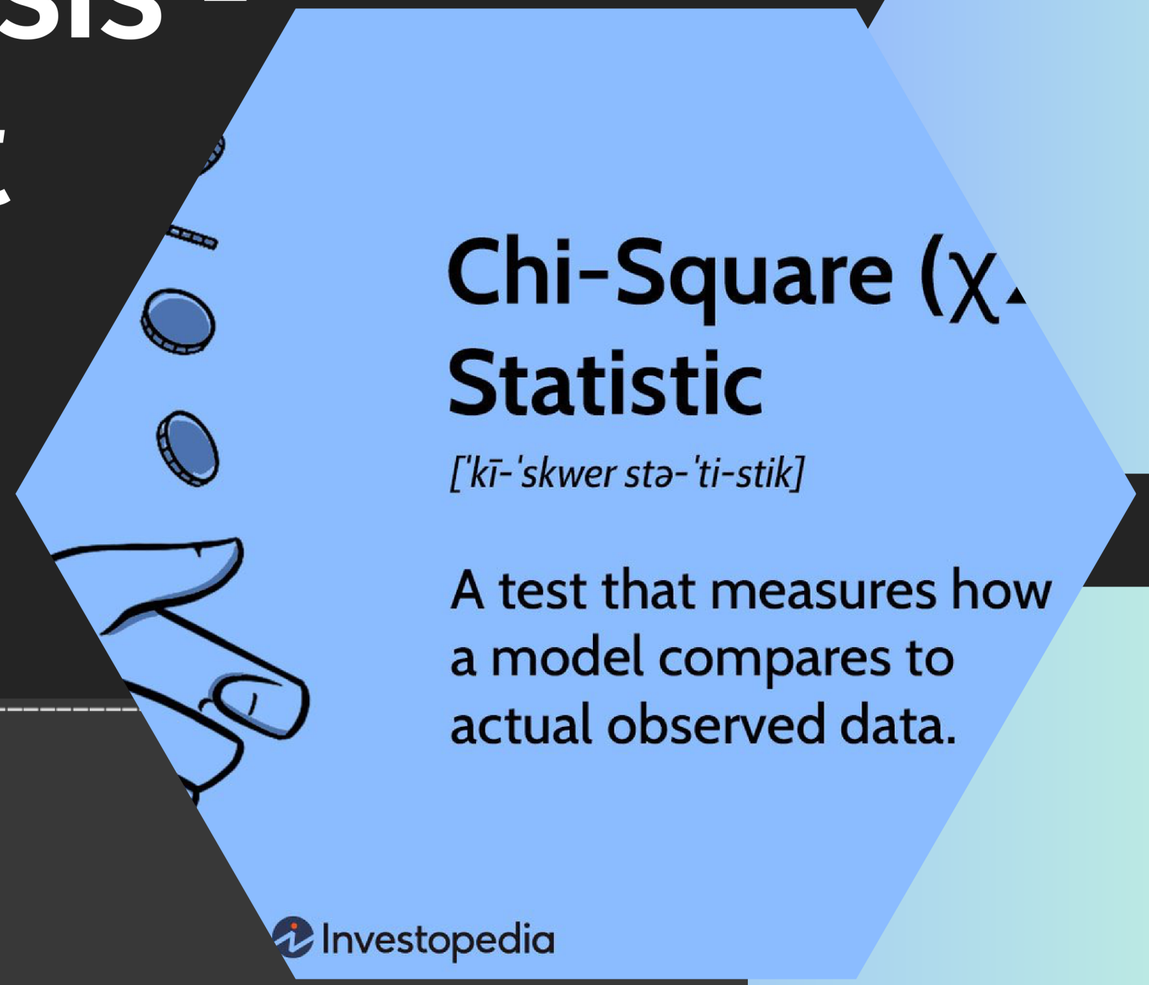


	PC0	PC1	PC2	PC3	PC4	PC5	PC6	PC7	PC8	PC9	PC10	PC11	PC12	PC13	PC14	PC15	PC16	PC17	PC18	PC19	PC20	PC21	PC22	PC23	PC24	PC25	PC26	PC27	PC28	PC29	PC30
0	3.06	5.49	-1.80	1.49	0.08	-3.36	0.49	-0.34	1.53	2.58	0.41	-0.92	0.58	0.41	-0.47	-1.07	0.07	0.11	0.31	-0.14	-0.28	-0.03	0.13	0.16	-0.35	-1.10	-0.64	-1.14	-0.05	1.88	0.07
1	-0.63	-1.19	-2.42	0.11	-0.15	0.87	-0.25	0.99	-2.55	-0.72	1.71	1.38	-1.01	-0.13	-0.59	-0.44	-0.11	-0.05	0.16	0.64	-1.31	-0.18	0.33	0.30	-0.05	0.86	-1.23	-0.38	-0.18	-0.58	-0.43
2	-0.42	-0.00	0.06	-0.73	0.07	0.56	-0.14	-1.05	-2.77	0.82	1.40	-1.70	-1.66	-0.07	-0.36	-0.67	0.48	0.26	-0.20	-1.31	-1.71	-0.07	0.25	1.00	2.18	-0.05	-0.97	-0.97	-0.74	-0.40	-0.70
3	-1.09	-0.94	1.50	-0.03	0.94	-0.36	-0.20	0.66	-0.23	0.42	-0.33	-2.13	0.90	0.21	-0.37	-0.67	0.26	0.05	0.10	-0.01	0.35	0.04	-0.38	0.35	1.47	-0.75	-0.11	-0.92	-0.75	0.60	0.26
4	-3.46	1.25	-1.56	0.53	1.19	-2.11	0.16	-0.68	-1.04	1.02	0.16	0.15	0.26	0.62	-1.17	-0.66	-0.03	-0.11	0.25	0.46	1.36	0.13	-0.11	-0.32	-1.14	-0.10	-0.20	-0.19	-0.35	-0.15	1.05
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
84081	2.43	-0.35	2.77	-0.72	-0.76	-0.94	0.32	-0.81	0.24	0.40	0.56	-0.19	0.11	-0.01	0.03	0.04	-0.01	-0.06	0.02	0.19	0.64	0.10	0.00	0.64	-0.46	-0.55	0.69	0.34	0.88	-0.04	0.05
84082	-0.06	0.73	1.82	-1.65	-0.00	-0.07	0.06	0.89	-0.33	1.19	1.24	-0.44	-0.36	0.88	-0.62	-0.71	-0.39	-0.47	0.69	2.17	0.85	0.28	-0.28	-0.44	1.29	-0.22	0.93	-0.04	-0.08	0.19	-0.85
84083	2.96	-1.62	-1.11	-2.71	2.63	-0.57	-0.01	-0.66	-2.49	-0.27	3.02	-3.46	-0.40	0.49	-0.72	-0.07	0.74	-0.15	-0.25	-2.30	2.98	0.27	-0.26	-0.51	3.36	-0.34	-0.39	-1.34	0.75	0.71	-0.65
84084	3.39	-5.13	-2.03	2.34	0.89	1.75	-0.84	0.64	-0.84	1.10	-0.03	1.51	3.27	-0.74	-1.04	-0.91	0.47	0.31	0.36	0.23	0.18	0.02	-1.83	-1.92	0.98	1.48	1.96	-2.10	0.23	0.13	-1.03
84085	-1.99	0.91	-2.67	-0.05	0.48	0.56	-0.14	-0.24	-1.55	-0.36	-0.07	-0.43	-0.82	0.48	0.17	0.34	-0.13	-0.15	-0.04	0.41	0.72	0.32	0.31	0.50	-1.41	0.13	-0.29	0.04	0.93	-0.49	-0.57

84086 rows × 31 columns

Correlation Analysis - Chi-Squared-Test

We see that between categorical columns and our target column loan status there are columns such as hardship_flag, verification_status_Source, purpose_sm, etc. which have no significant association with loan status




```
--------------------------------------------------------------------------------

Contingency table for addr_state_WV and Loan Status:
loan_status       0      1
addr_state_WV
0              7569  76004
1                49    464

Chi-squared test results for addr_state_WV and Loan Status:
chi2 = 0.09745792886003712
p-value = 0.7549018120284605
dof = 1

There is no significant association between addr_state_WV and Loan Status with p-value of 0.7549018120284605.

--------------------------------------------------------------------------------
```

MODEL EXPLORATION AND MODEL SELECTION

LazyClassifier is a Python library that allows us to quickly build and test several machine learning models without requiring a lot of code.
It is called "lazy" because it automates the process of model building and evaluation, allowing us to quickly test a range of models and compare their performance.

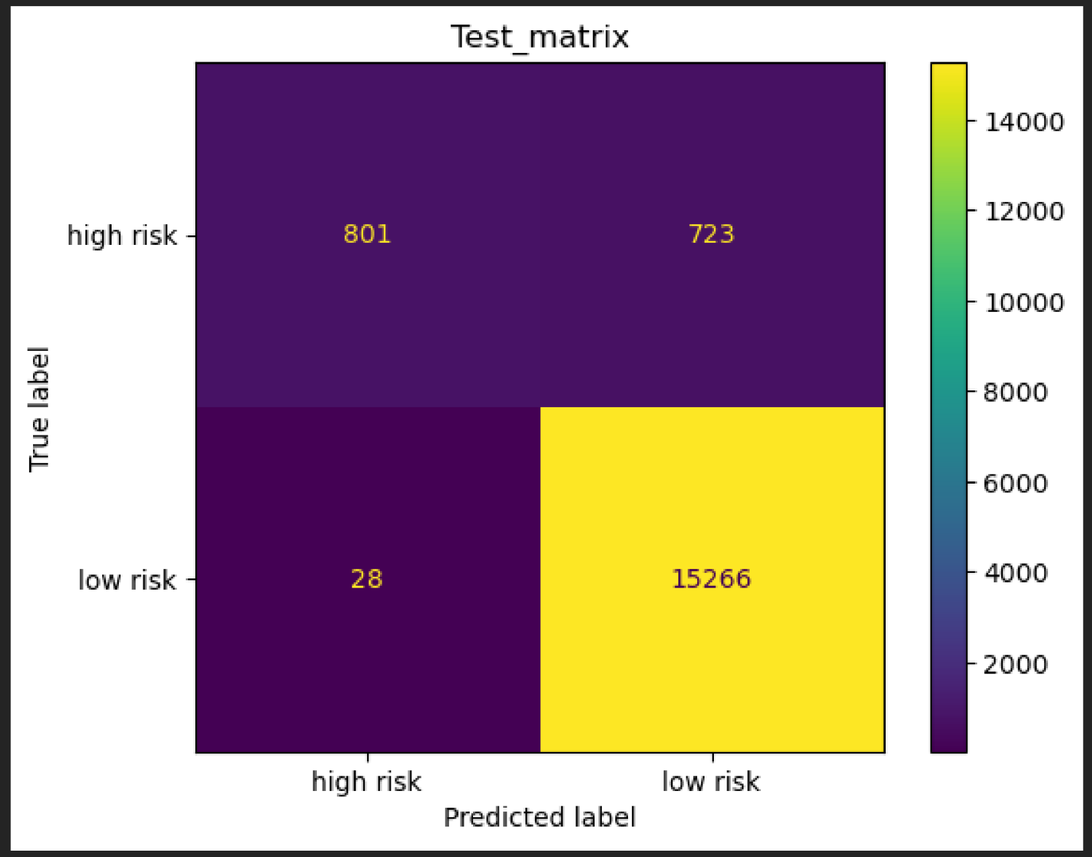
1.Linear SVC Classifier
2. Logistic Regression
3. Logical Regression with SMOTE over-sampling
4. XGB Classifier
5. Bagging Classifier
6. Decision Tree Classifier

LAZY CLASSIFIER OUTCOME

Model	Accuracy	Balanced Accuracy	ROC AUC	F1 Score	Time Taken
LogisticRegression	0.96	0.78	0.78	0.95	1.49
PassiveAggressiveClassifier	0.92	0.77	0.77	0.92	0.62
DecisionTreeClassifier	0.94	0.77	0.77	0.94	3.73
BaggingClassifier	0.96	0.77	0.77	0.95	23.13
SGDClassifier	0.95	0.77	0.77	0.95	1.06
Perceptron	0.94	0.77	0.77	0.94	0.51
LGBMClassifier	0.96	0.76	0.76	0.95	0.91
LinearSVC	0.96	0.76	0.76	0.95	20.71
CalibratedClassifierCV	0.96	0.76	0.76	0.95	75.69
BernoulliNB	0.91	0.76	0.76	0.91	0.42
AdaBoostClassifier	0.95	0.75	0.75	0.95	10.83
XGBClassifier	0.95	0.75	0.75	0.95	3.96
RandomForestClassifier	0.95	0.75	0.75	0.95	20.13
QuadraticDiscriminantAnalysis	0.95	0.73	0.73	0.94	0.70
NearestCentroid	0.81	0.73	0.73	0.84	0.35
SVC	0.95	0.71	0.71	0.94	543.71
LinearDiscriminantAnalysis	0.94	0.70	0.70	0.93	1.10
ExtraTreesClassifier	0.94	0.69	0.69	0.93	15.54
ExtraTreeClassifier	0.90	0.69	0.69	0.90	0.47
RidgeClassifier	0.93	0.64	0.64	0.92	0.52
RidgeClassifierCV	0.93	0.64	0.64	0.92	0.83
KNeighborsClassifier	0.92	0.57	0.57	0.89	2.58
GaussianNB	0.09	0.50	0.50	0.02	0.43
DummyClassifier	0.91	0.50	0.50	0.87	0.29

Credrisk: IE 7275 Data Mining in Engineering

Performance evaluation for Linear SVC classifier



Test set Performance:
Balanced Accuracy Score 0.7618798839336922
Accuracy Score: 0.955354631941967
Precision Score: 0.9558183406687925
Recall Score: 0.955354631941967
F1 Score: 0.9492467457951946

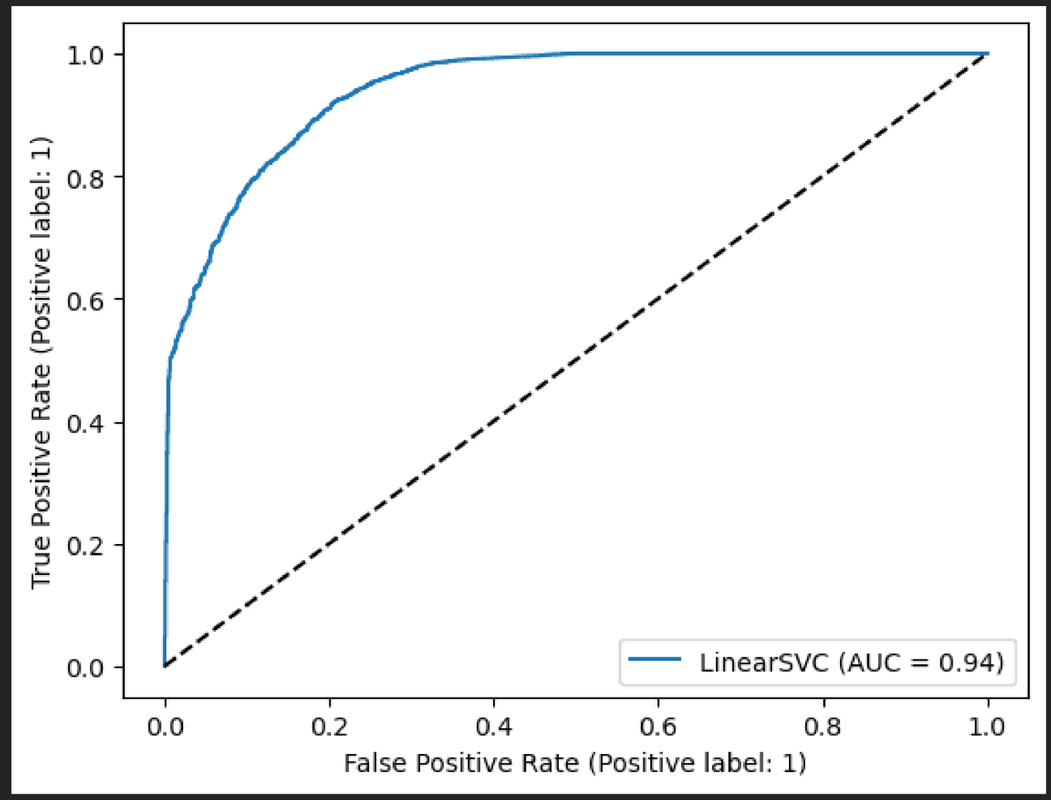
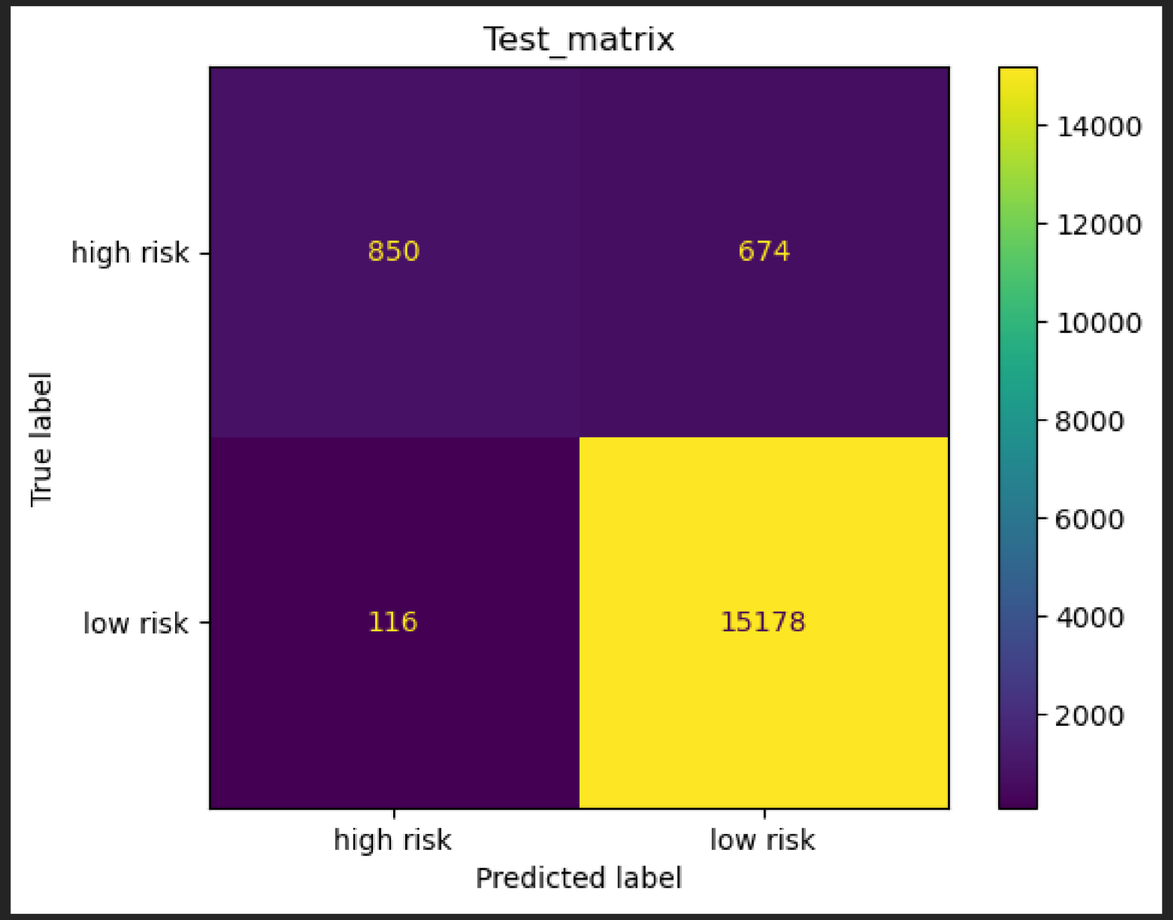


Fig 5.1.3: ROC and AUC Curve Linear SVC Classifier:

Performance evaluation for Logistic Regression

Test set Performance:
Balanced Accuracy Score 0.7750790542119858
Accuracy Score: 0.953026519205613
Precision Score: 0.9504530263405578
Recall Score: 0.953026519205613
F1 Score: 0.948184005165446



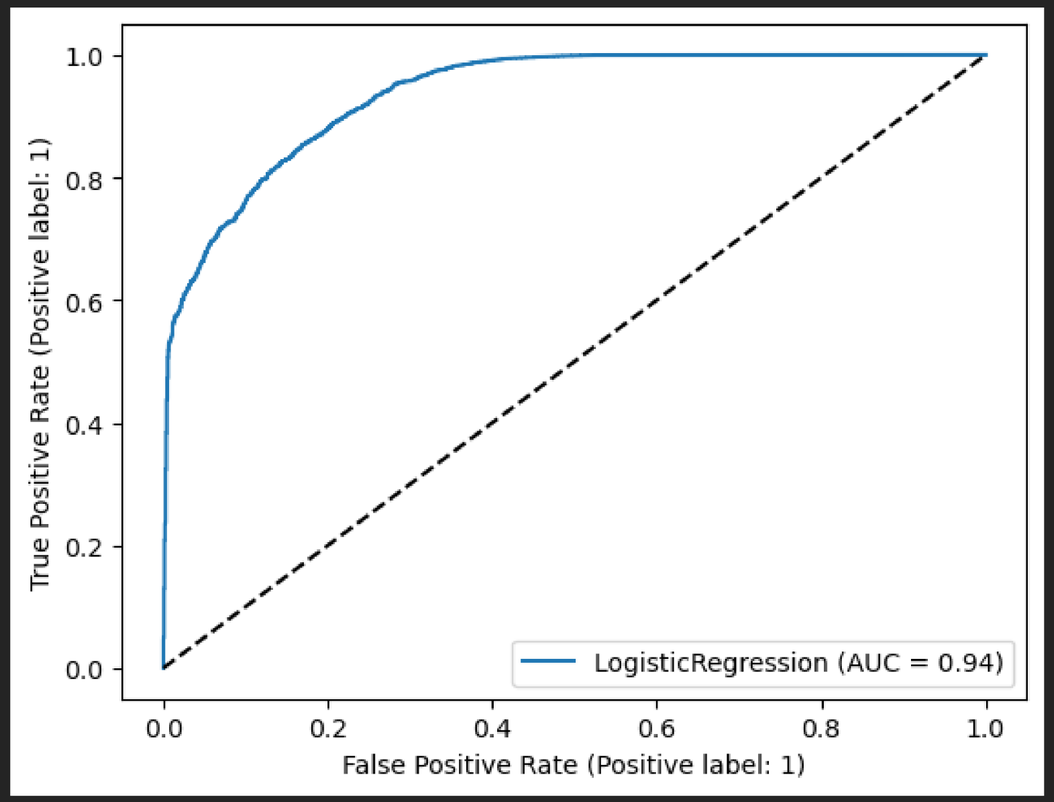
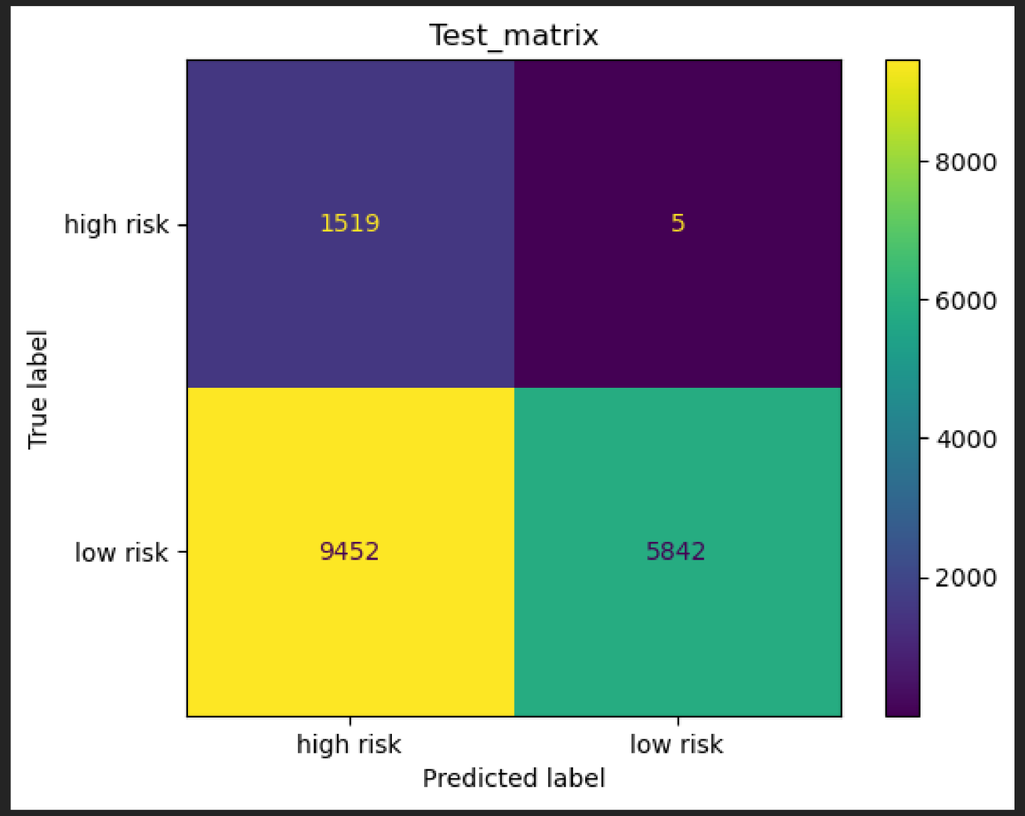


Fig 5.1.3: ROC and AUC Curve

Performance evaluation for Logical Regression with SMOTE over-sampling

Test set Performance:
Balanced Accuracy Score 0.6911649917790064
Accuracy Score: 0.4376858128195980 6
Precision Score: 0.9211516431447702
Recall Score: 0.4376858128195980 6
F1 Score: 0.5246211624552783



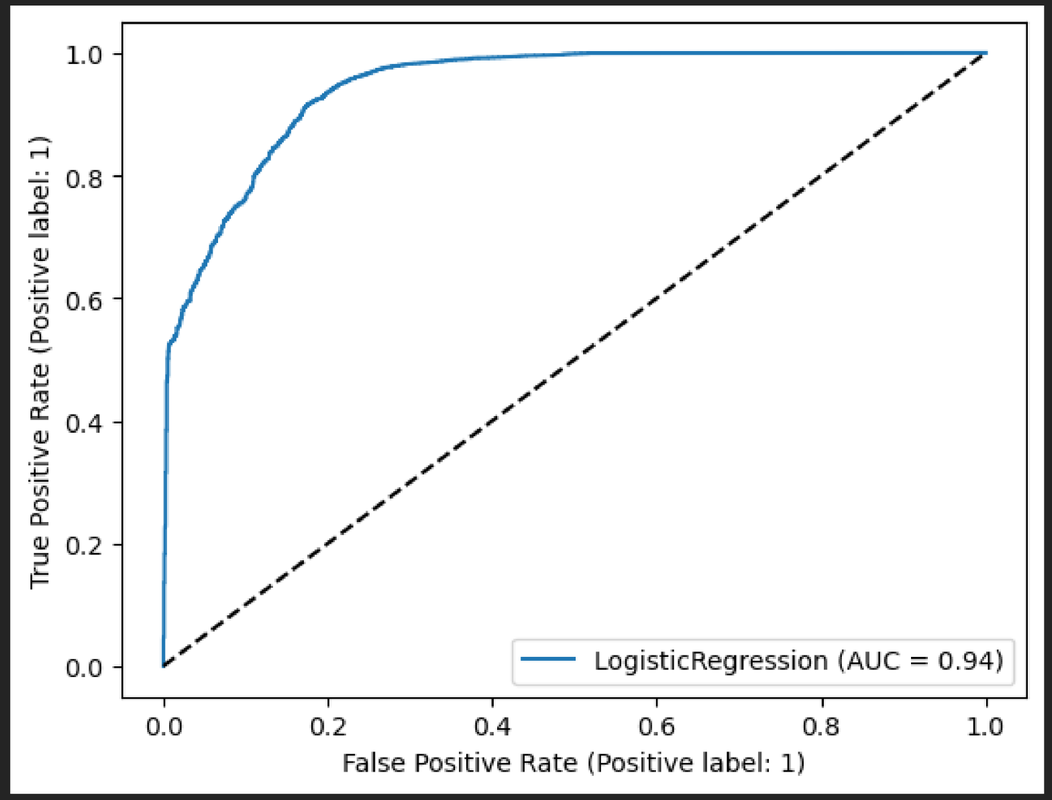
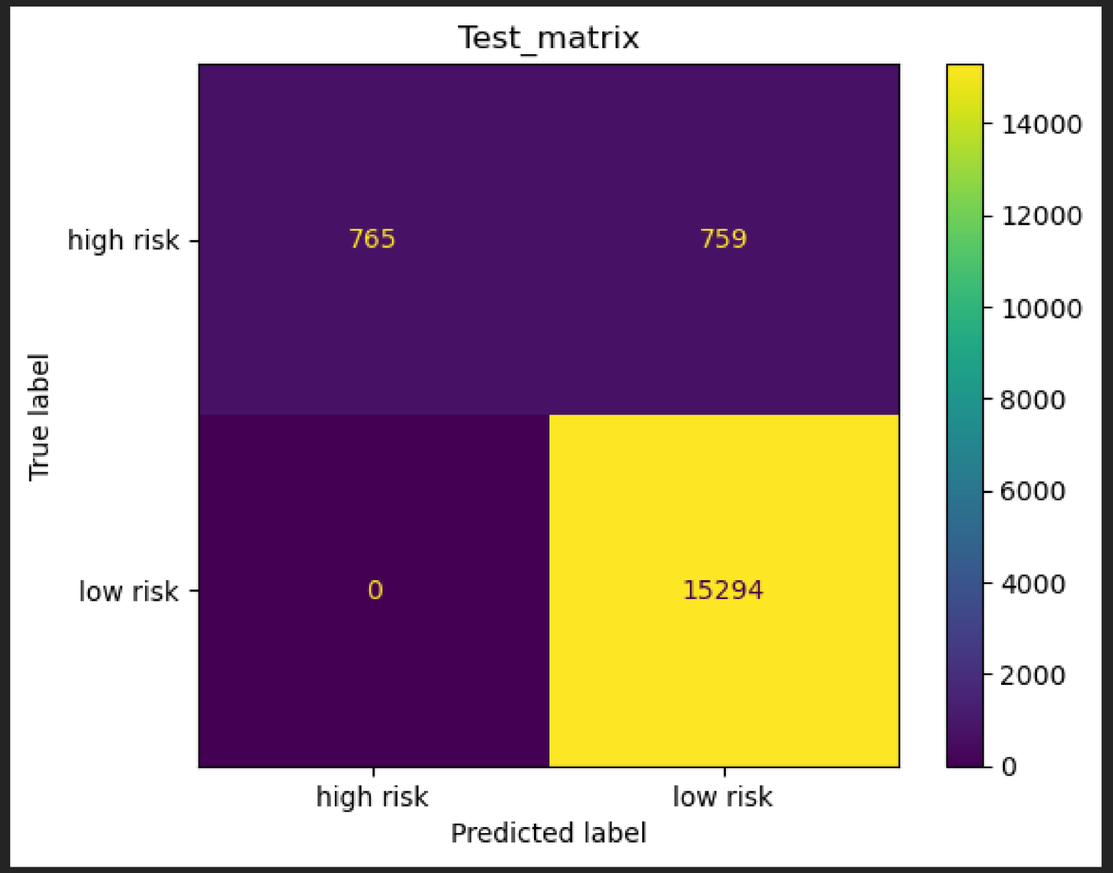


Fig 5.1.3: ROC and AUC Curve

Performance evaluation for XGB Classifier

Test set Performance:
Balanced Accuracy Score 0.7509842519685039
Accuracy Score: 0.9548697823760257
Precision Score: 0.9570035788736646
Recall Score: 0.9548697823760257
F1 Score: 0.9479338771397345



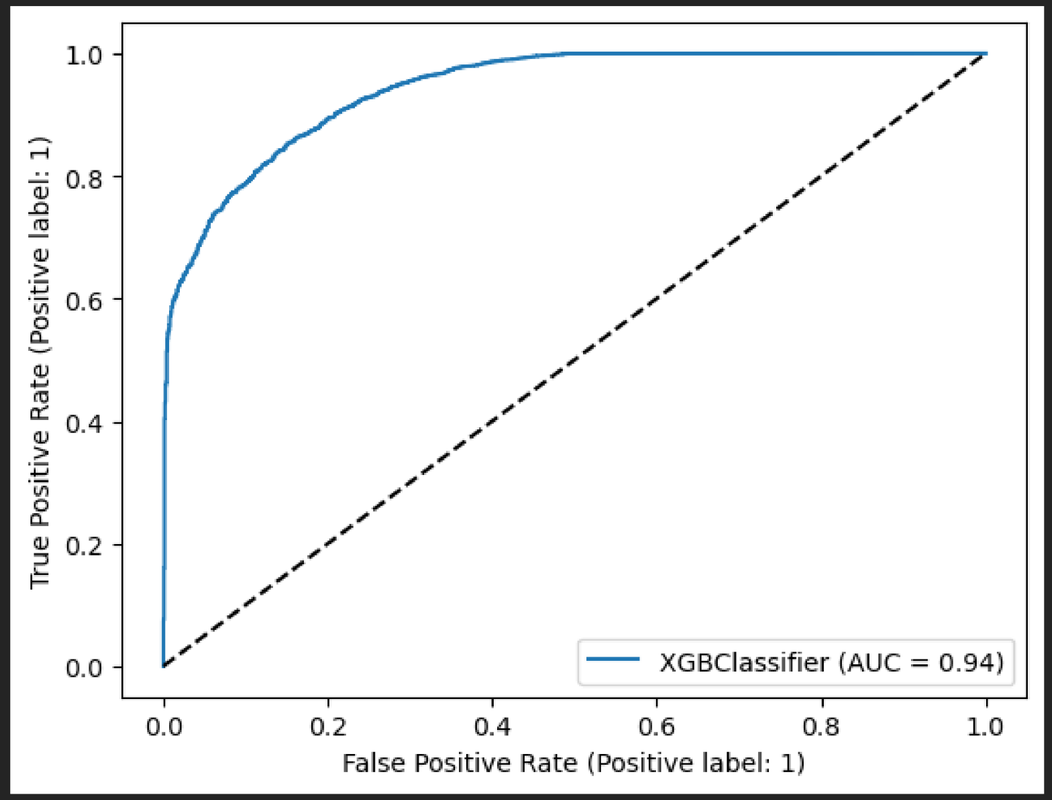
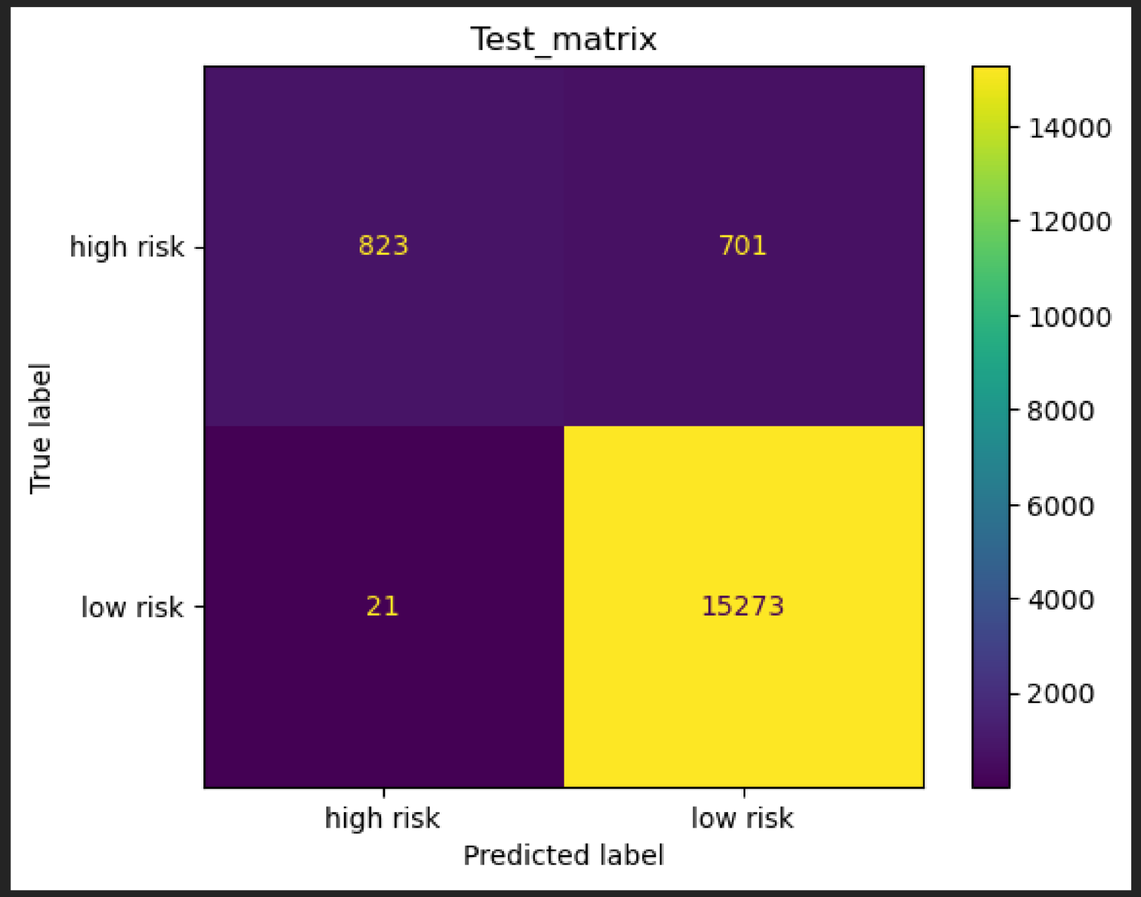


Fig 5.1.3: ROC and AUC Curve

Performance evaluation for Bagging Classifier

Test set Performance:
Balanced Accuracy Score 0.769326579616936
Accuracy Score: 0.9570698061600666
Precision Score: 0.9578381234588981
Recall Score: 0.9570698061600666
F1 Score: 0.9513726541429909



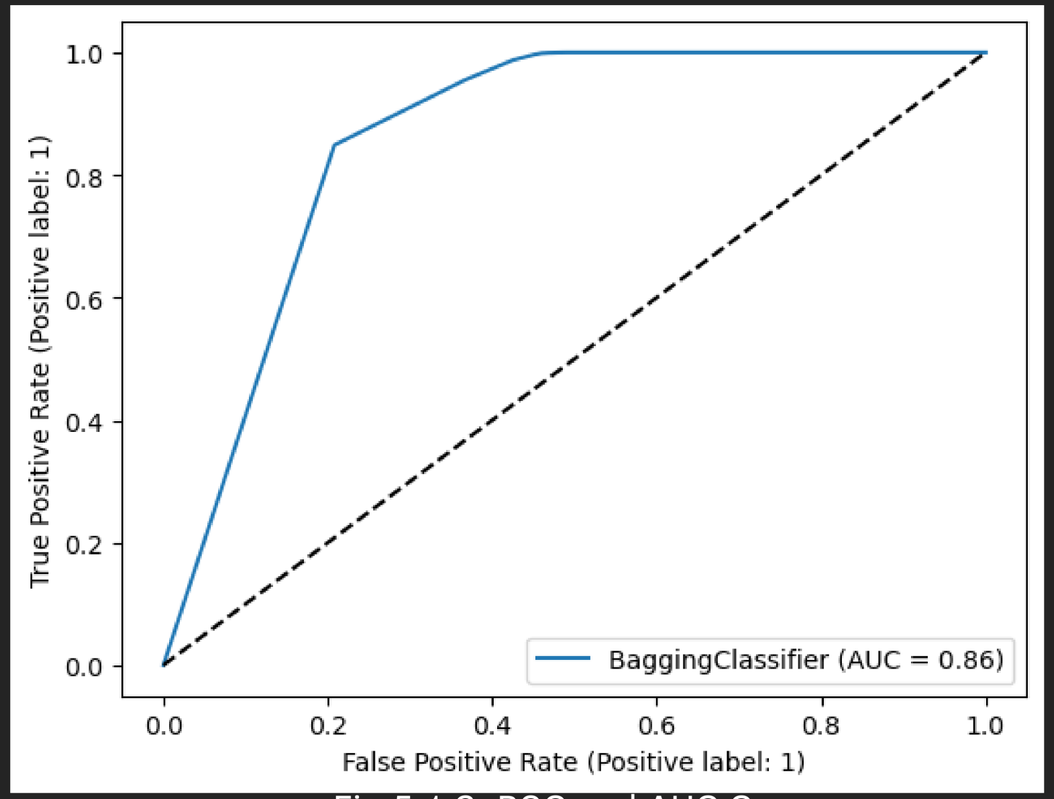
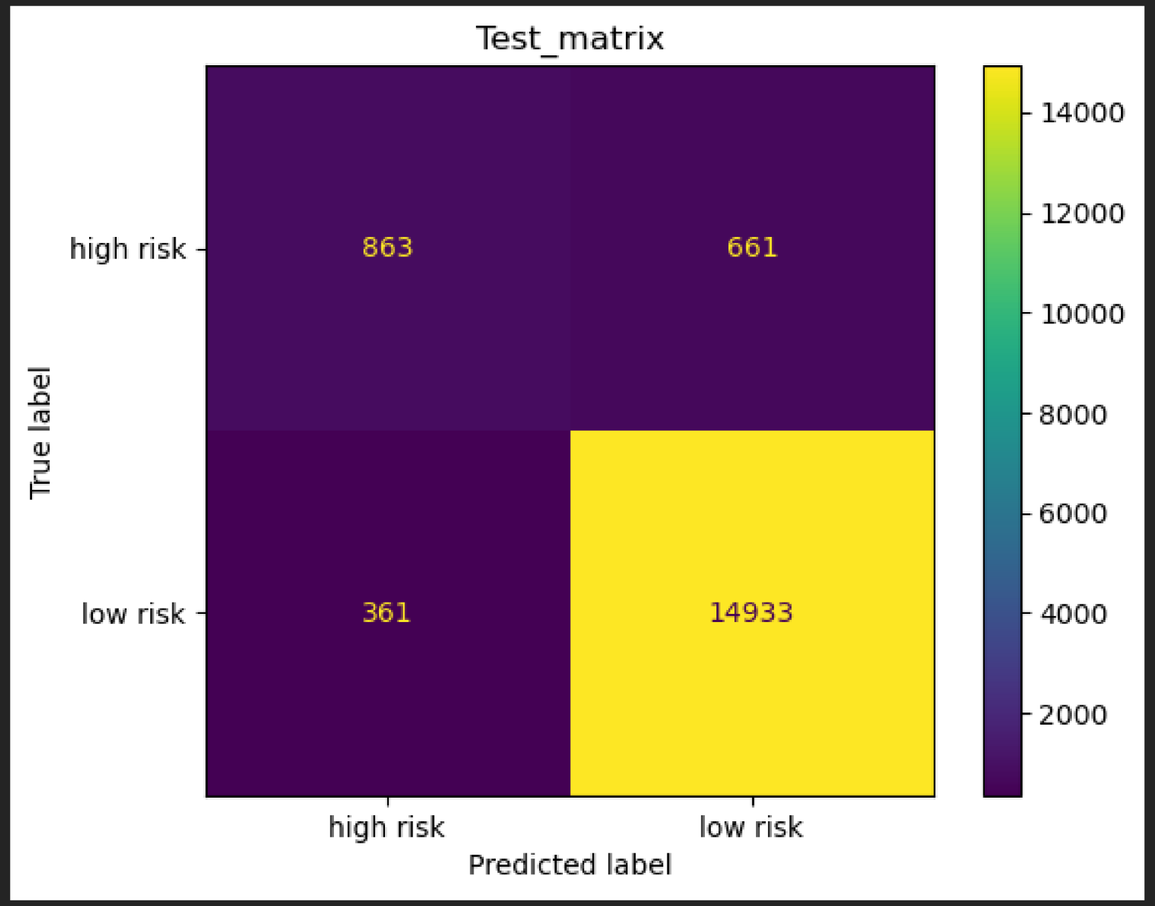


Fig 5.1.3: ROC and AUC Curve

Performance evaluation for Decision Tree Classifier

Test set Performance:
Balanced Accuracy Score 0.771334469077987 5
Accuracy Score: 0.9392317754786538
Precision Score: 0.9347268425044035
Recall Score: 0.9392317754786538
F1 Score: 0.9362098386408739



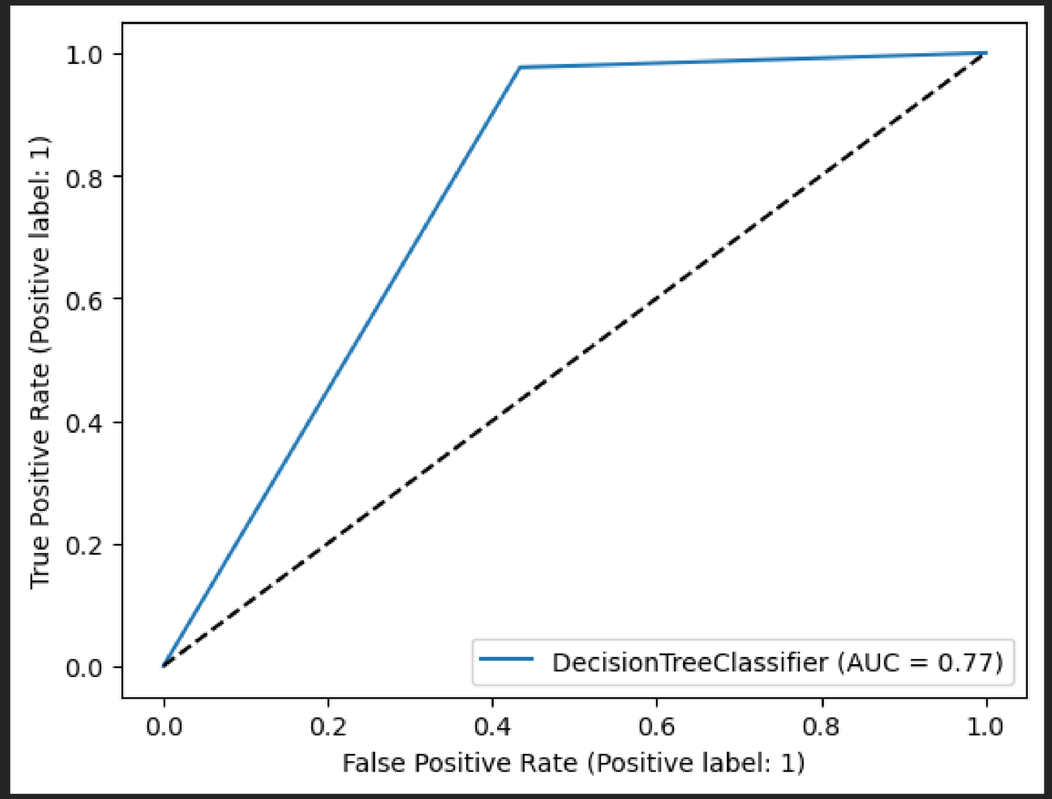
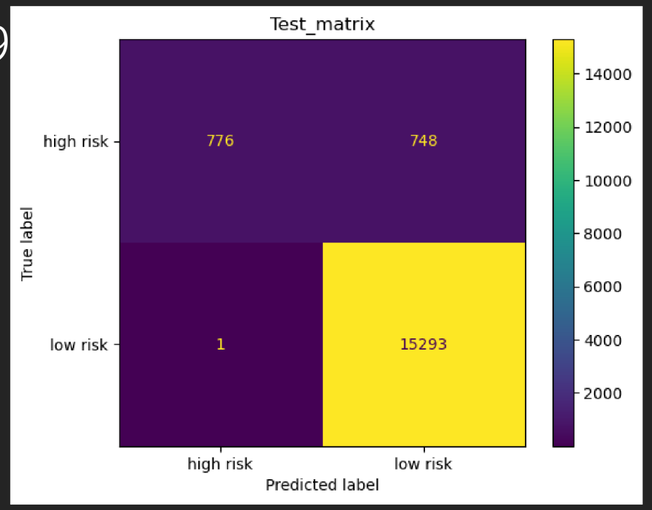


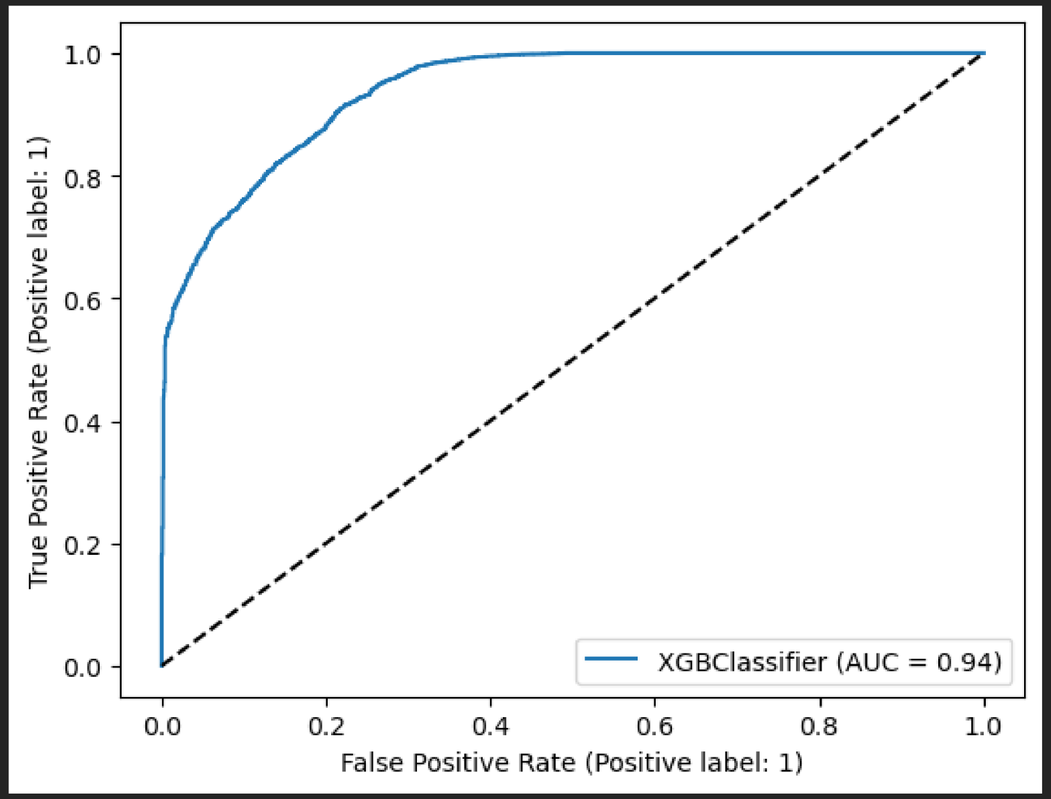
Fig 5.1.3: ROC and AUC Curve

Hyper Parameter Tunning for XGB Classifier

- Defined the range of hyperparameters.
- A grid-search was used to find the best model using learning_rate=0.1, n_estimators=100, colsample_bytree= 0.5, max_depth= 7, subsample= 1.0
- Balanced Accuracy Score 0.7545604832938448
- Accuracy Score: 0.9554643833987394
- Precision Score: 0.9574783922285974
- Recall Score: 0.9554643833987394
- F1 Score: 0.9487661753848690

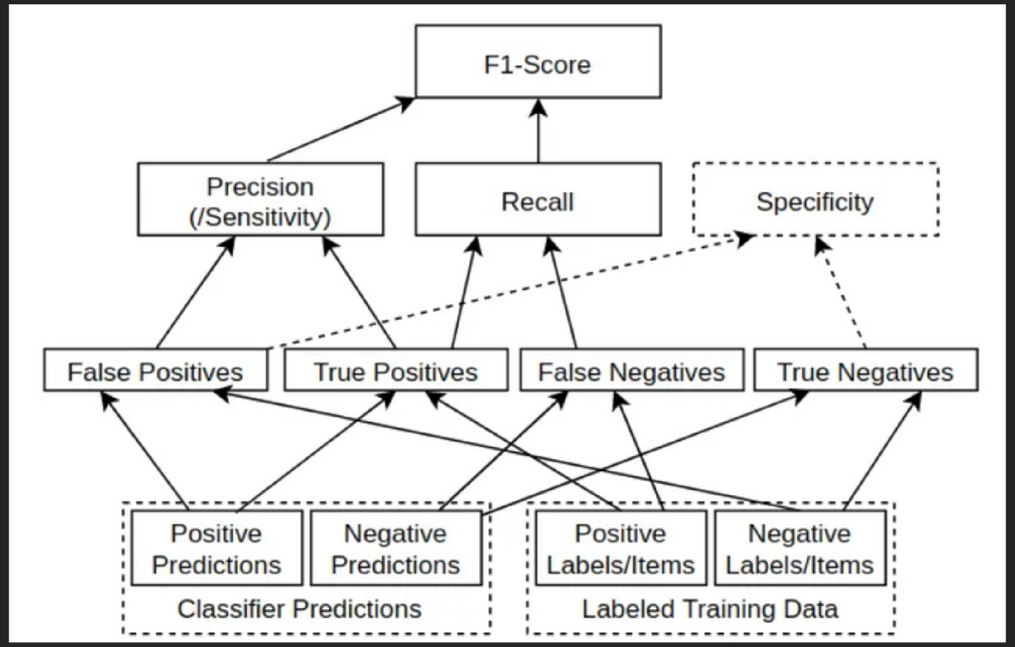
	precision	recall	f1-score	support
0	1.00	0.51	0.67	1524
1	0.95	1.00	0.98	15294
accuracy			0.96	16818
macro avg	0.98	0.75	0.82	16818
weighted avg	0.96	0.96	0.95	16818





Model Performance comparison

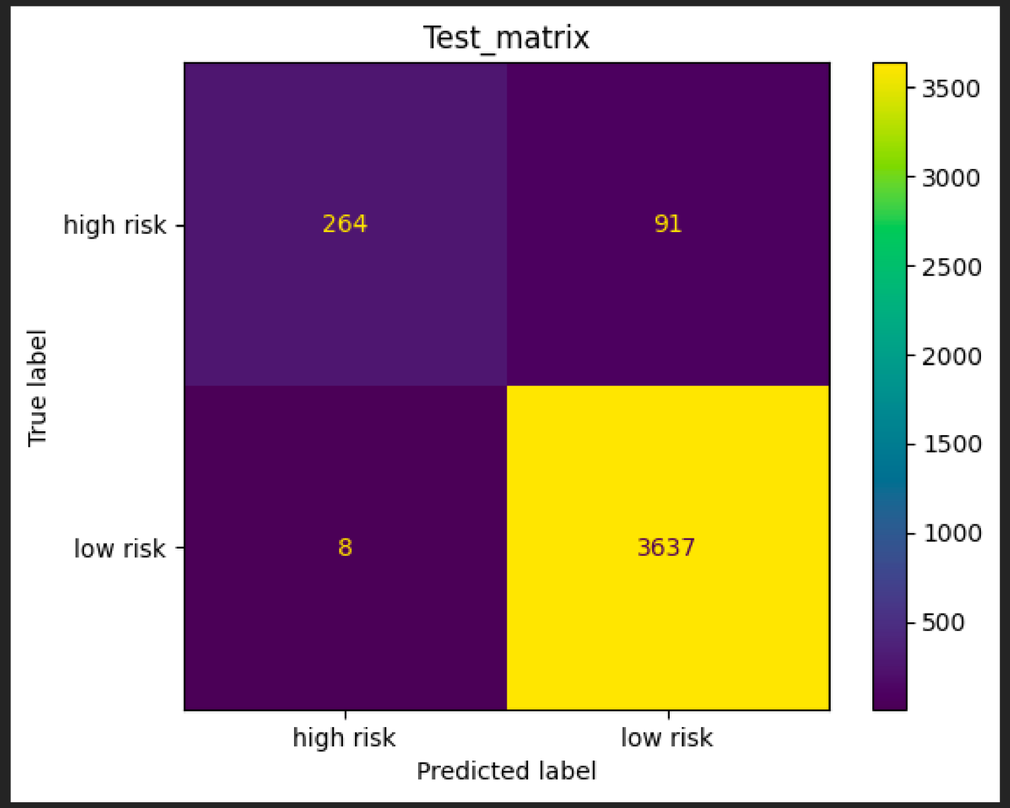
Models	Balanced Accuracy	Precision0	Precision1	Recall0	Recall1	F1-score0	F1-score1
Linear SVC Classifier	0.76	0.97	0.95	0.53	1	0.68	0.98
Logistic Regression	0.77	0.88	0.96	0.56	0.99	0.68	0.97
Logistic Regression with SMOTE	0.69	0.14	1	1	0.38	0.24	0.55
XGB Classifier	0.76	1	0.95	0.50	1	0.67	0.98
Bagging Classifier	0.76	0.98	0.96	0.54	1	0.70	0.98
Decision Tree Classifier	0.77	0.71	0.96	0.57	0.98	0.63	0.97

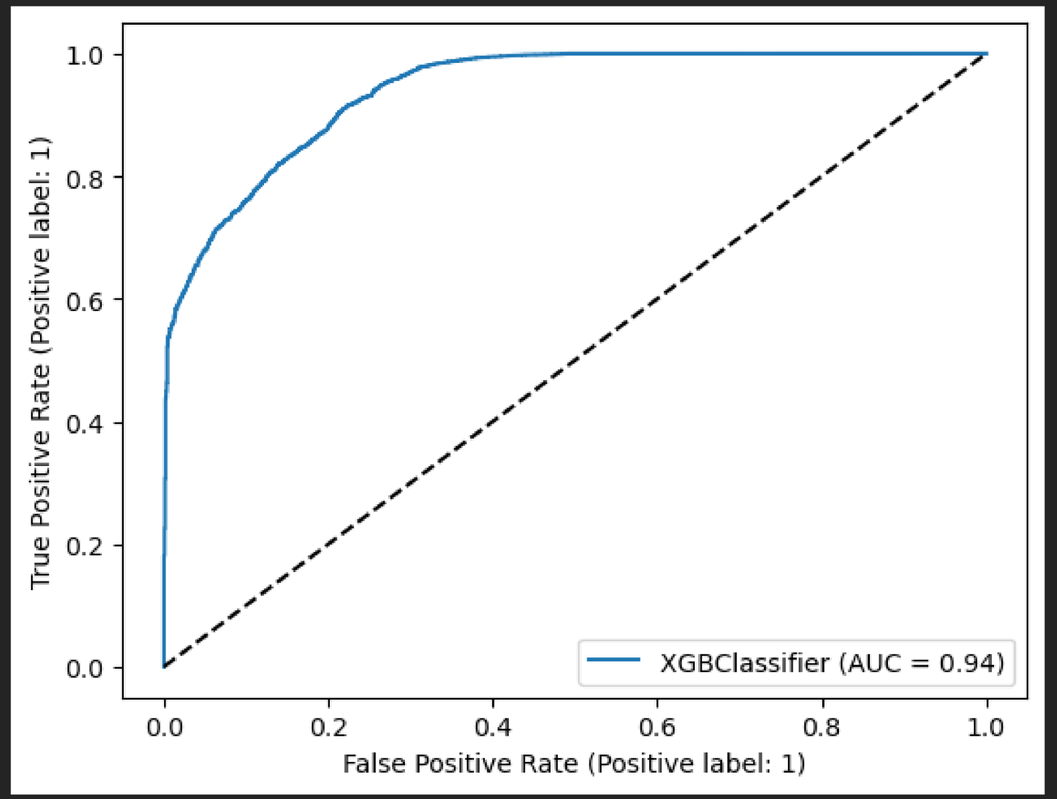


The F1-score metric is used as the assessment criterion as we need ‘True positive’ to be the target evaluation metric, leading to the combination of Precision(/sensitivity) and recall.

Performance visualization of the selected model

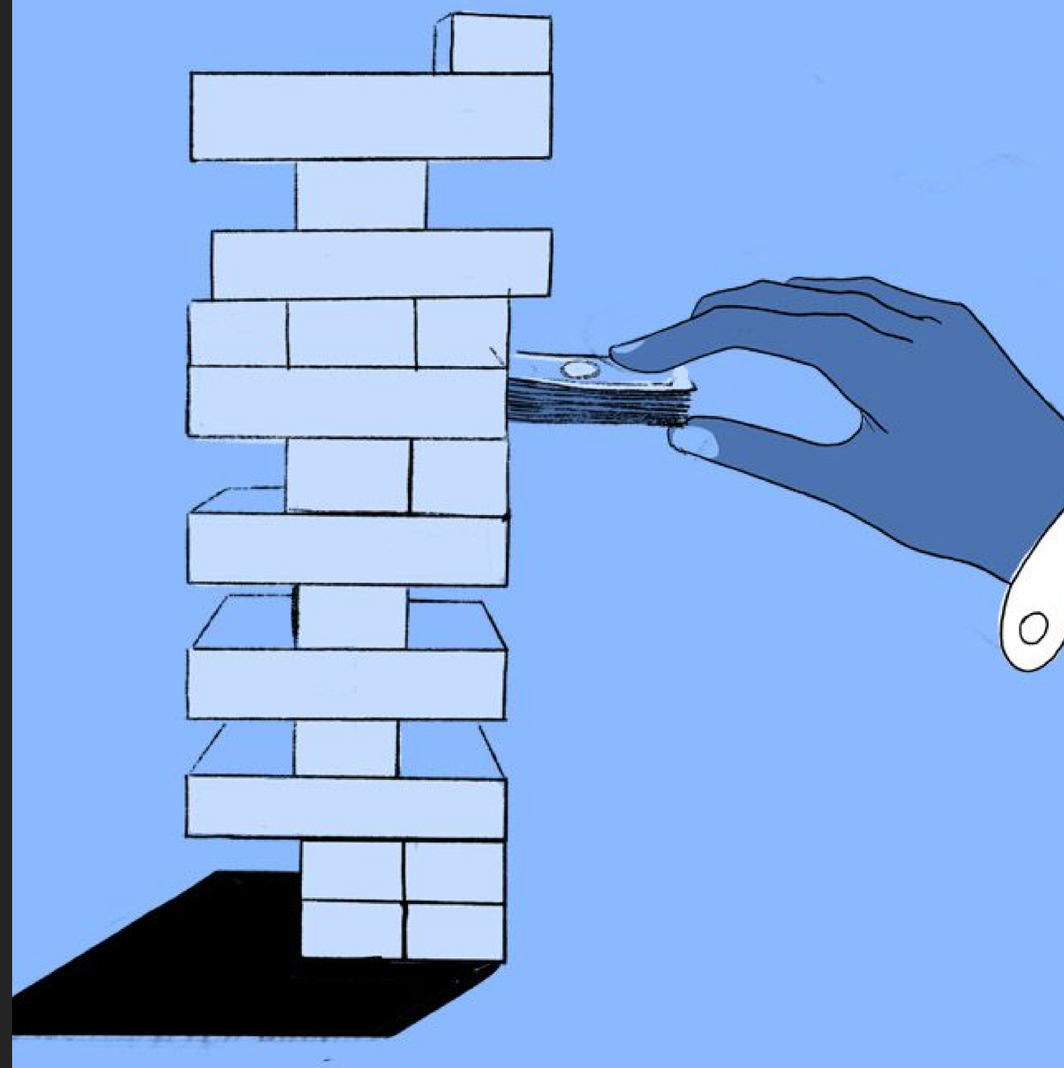
For the High_risk(0class) class, the XGB Classifier had the greatest F1-score of 0.84, demonstrating a superior balance between accuracy(0.87), precision(0.95) and recall(0.74). The model earned a balanced accuracy of 0.87, which is a positive sign for the model's overall performance. The XGB Classifier is therefore the most suitable option for this purpose.



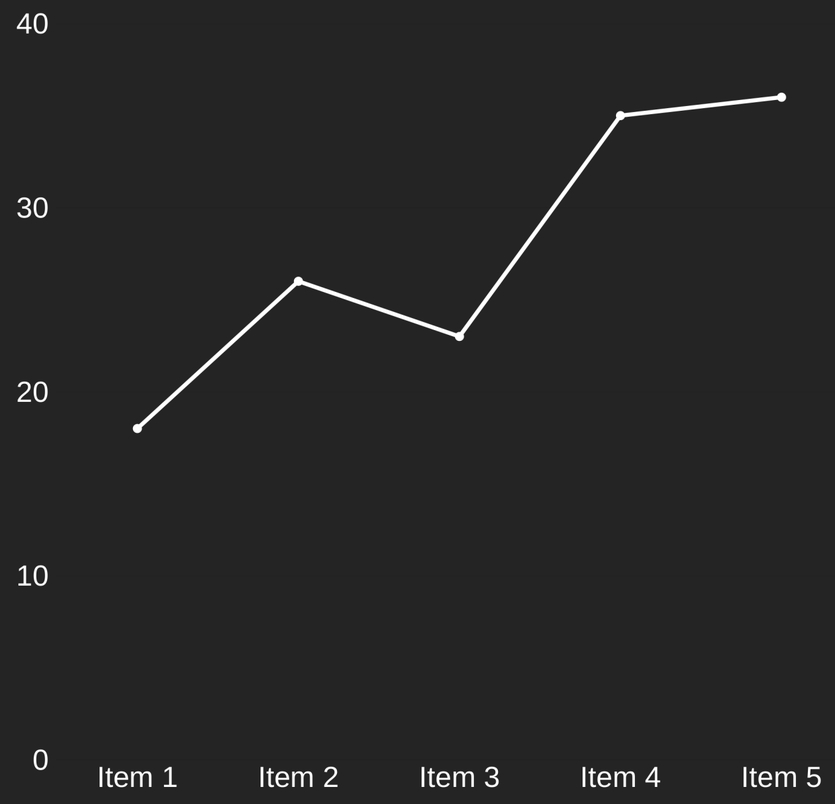


Limitations

1. LIMITED DATASET: THE DATASET USED IN THIS PROJECT ONLY INCLUDES A SMALL PORTION OF ALL LOAN APPLICATIONS MADE IN THE MARKET. THEREFORE, THE FINDINGS MAY NOT BE REPRESENTATIVE OF THE ENTIRE POPULATION.
2. MISSING DATA: THE DATASET CONTAINS SOME MISSING DATA, WHICH MAY HAVE AFFECTED THE ACCURACY OF THE RESULTS.
3. LIMITED FEATURE SELECTION: ALTHOUGH SOME FEATURE SELECTION TECHNIQUES WERE USED TO SELECT THE MOST RELEVANT FEATURES, THERE MAY BE OTHER IMPORTANT FEATURES THAT WERE NOT INCLUDED IN THE ANALYSIS.
4. TIMEFRAME OF DATA COLLECTION: THE DATA WAS COLLECTED OVER A SPECIFIC TIME PERIOD, AND ECONOMIC CIRCUMSTANCES MAY HAVE CHANGED SINCE THEN. THIS COULD LIMIT THE MODEL'S ABILITY TO PREDICT LOAN DEFAULTS IN CURRENT ECONOMIC SITUATIONS.



FUTURE SCOPE



1. Increase data diversity: To improve the model's accuracy and robustness, it is essential to incorporate additional variables and data sources to expand the diversity of loan applications considered.
2. Real-time data: Collecting real-time data on economic conditions and borrower behavior could enhance the model's performance and ability to predict loan defaults accurately.
3. Advanced modeling techniques: The use of more advanced machine learning algorithms and techniques, such as deep learning, may provide superior predictive power and accuracy.
4. Deployment of the model: The predictive model developed in this project can be integrated into a loan provider's decision-making process to automate the loan approval process and enhance risk management practices with a suitable user interface.

THANK YOU