# **Home Credit Default Risk: Model Evaluation Report**

## **1. Introduction**

Home Credit aims to provide financial inclusion to individuals who lack access to traditional credit by using alternative data sources such as telco and transaction history to assess borrowers' ability to repay loans. Accurate predictions of loan defaults are critical to ensuring that creditworthy individuals are not rejected and that loans are structured to avoid over-indebtedness.

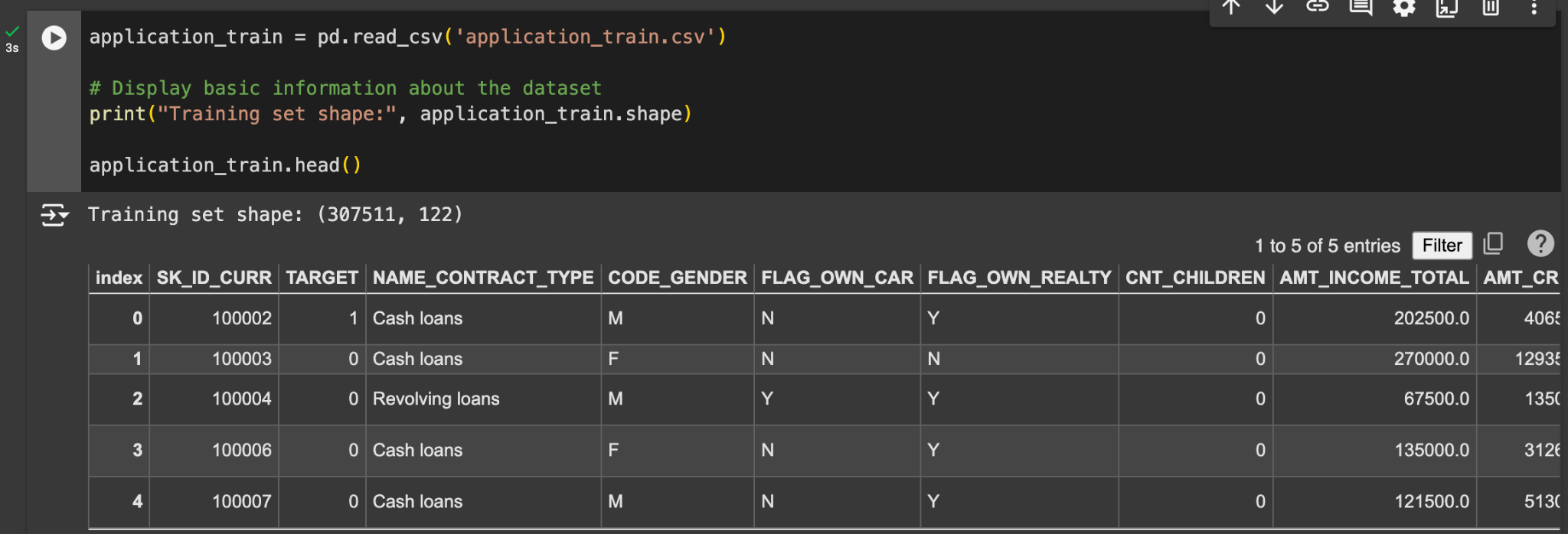
This report evaluates the performance of multiple machine learning models in predicting loan defaults for **Home Credit**. The models tested are:

* Random Forest
* Logistic Regression
* K-Nearest Neighbors (KNN)
* Linear Support Vector Machine (SVM)
* Neural Networks

## **2. Dataset Overview**

The dataset used for this competition contains information on loan applicants with 121 columns, including:

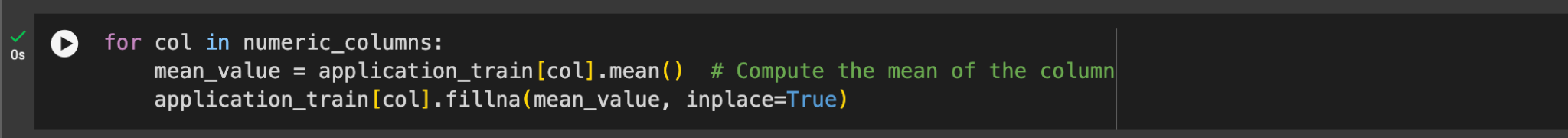
* **Application Data**: Features like applicant age, income, and employment status.
* **Previous Loan Data**: Data about any past loans, including loan amounts, repayment history, and loan status.
* **Telco Data**: Mobile usage data from telecommunications providers, which serves as an alternative indicator of financial behaviour.
* **Transaction Data**: Records of applicants' financial transactions, which provide insights into their spending and savings habits.

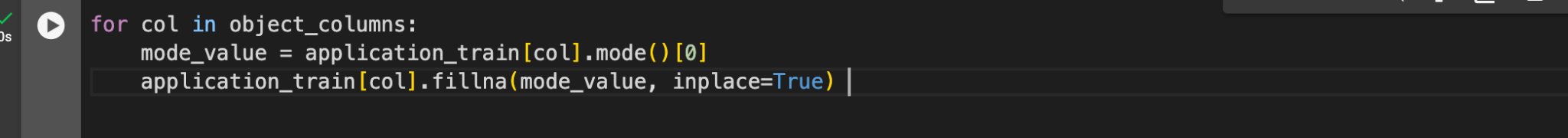


### 

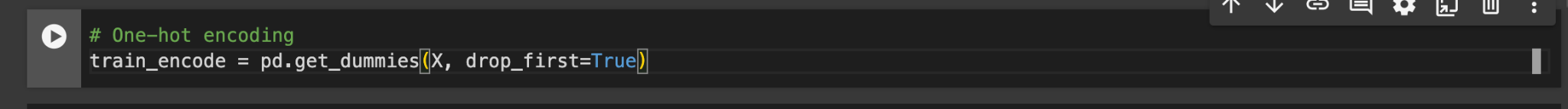
### **Data Preprocessing:**

* **Handling Missing Data**: Missing values were imputed using median values for numerical columns and the mode for categorical columns.

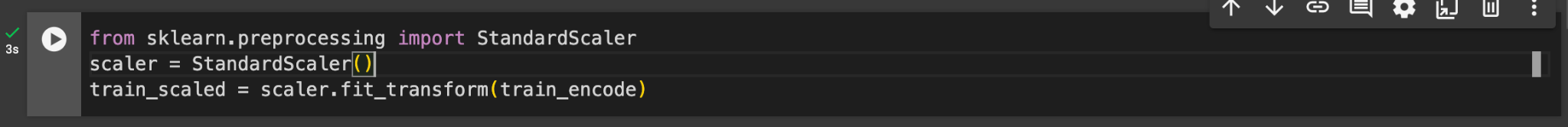




* **Feature Encoding**: Categorical variables were encoded using one-hot encoding to convert them into a suitable format for the machine learning models.



* **Feature Scaling**: Features were standardised (zero mean, unit variance) to ensure equal contribution from all features, especially for models like **KNN** and **SVM**, which are sensitive to the scale of the data.



## **3. Methodology**

### **Model Selection**

The following models were selected and evaluated:

1. **Random Forest**: An ensemble learning method based on decision trees that is robust to overfitting and effective at capturing non-linear relationships in the data. It works well with a mix of numerical and categorical features.
2. **Logistic Regression**: A simple, interpretable linear model commonly used for binary classification. It is computationally efficient and provides a good baseline for comparison.
3. **K-Nearest Neighbors (KNN)**: A non-parametric method that classifies instances based on the majority class of their nearest neighbours. KNN is sensitive to the distance between points and requires feature scaling.
4. **Linear Support Vector Machine (SVM)**: A linear classifier that separates classes by finding the hyperplane with the largest margin. It is particularly effective for high-dimensional spaces.
5. **Neural Networks**: 6 fully connected layers (1 input,3 dense layers,batch normalization layer,output layer).Traditional models like Logistic Regression or even Random Forests might struggle to fully exploit these interactions without extensive feature engineering.

### **Evaluation Metrics:**

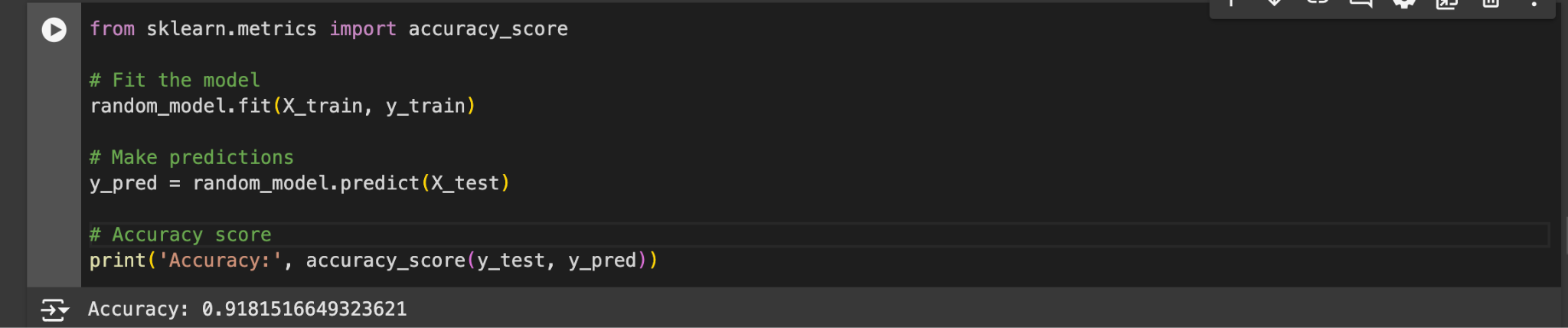
The following metrics were used to evaluate model performance:

* **Accuracy**: The overall proportion of correct predictions.
* **Precision**: The proportion of true positive predictions among all positive predictions.
* **Recall**: The proportion of true positive predictions among all actual positives (important in the context of loan defaults).
* **F1-Score**: The harmonic mean of precision and recall, useful for imbalanced datasets.
* **AUC-ROC**: The area under the Receiver Operating Characteristic curve, which measures the ability of the model to distinguish between the two classes.

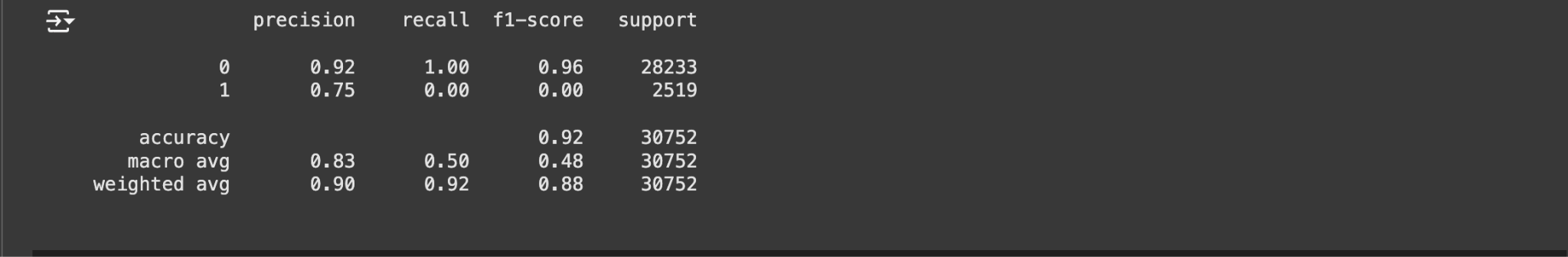
## **4. Results**

### **Models Performance**

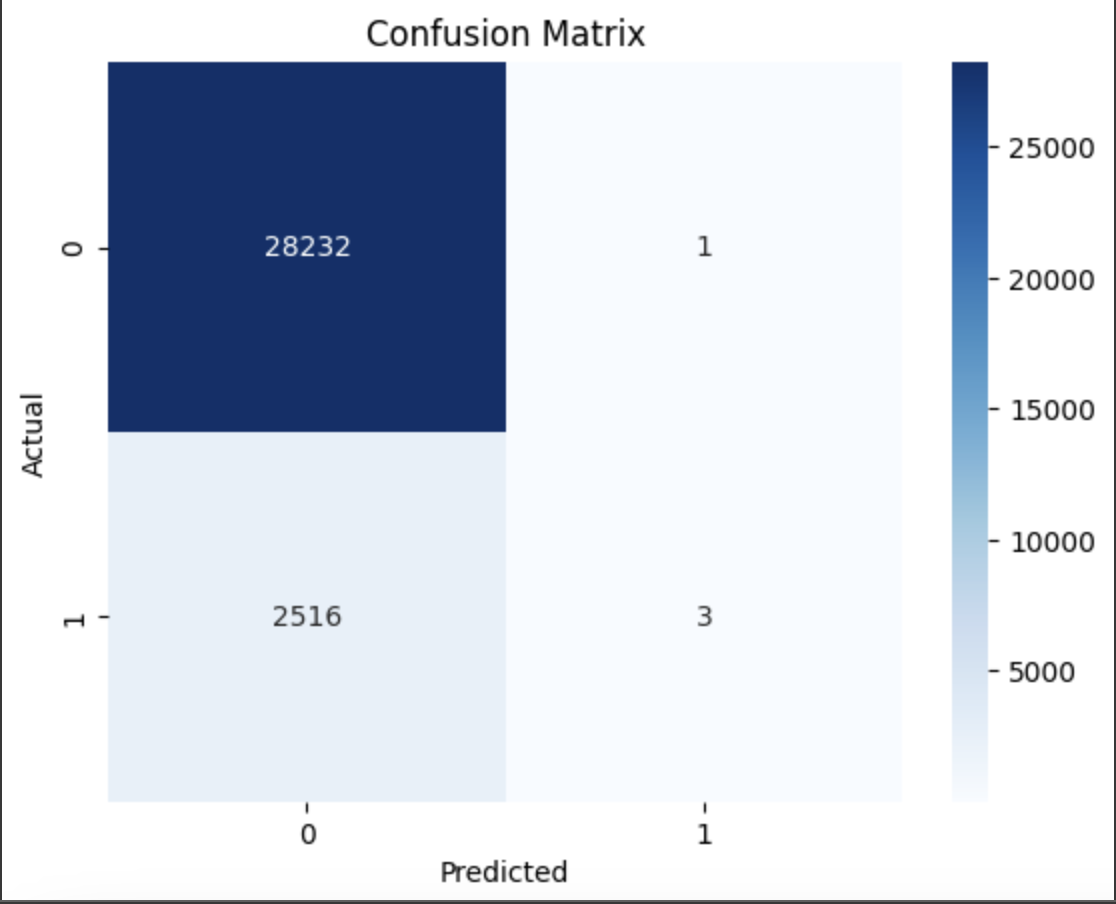
1. **Random Forest:** 
   1. **Accuracy Score:**

****

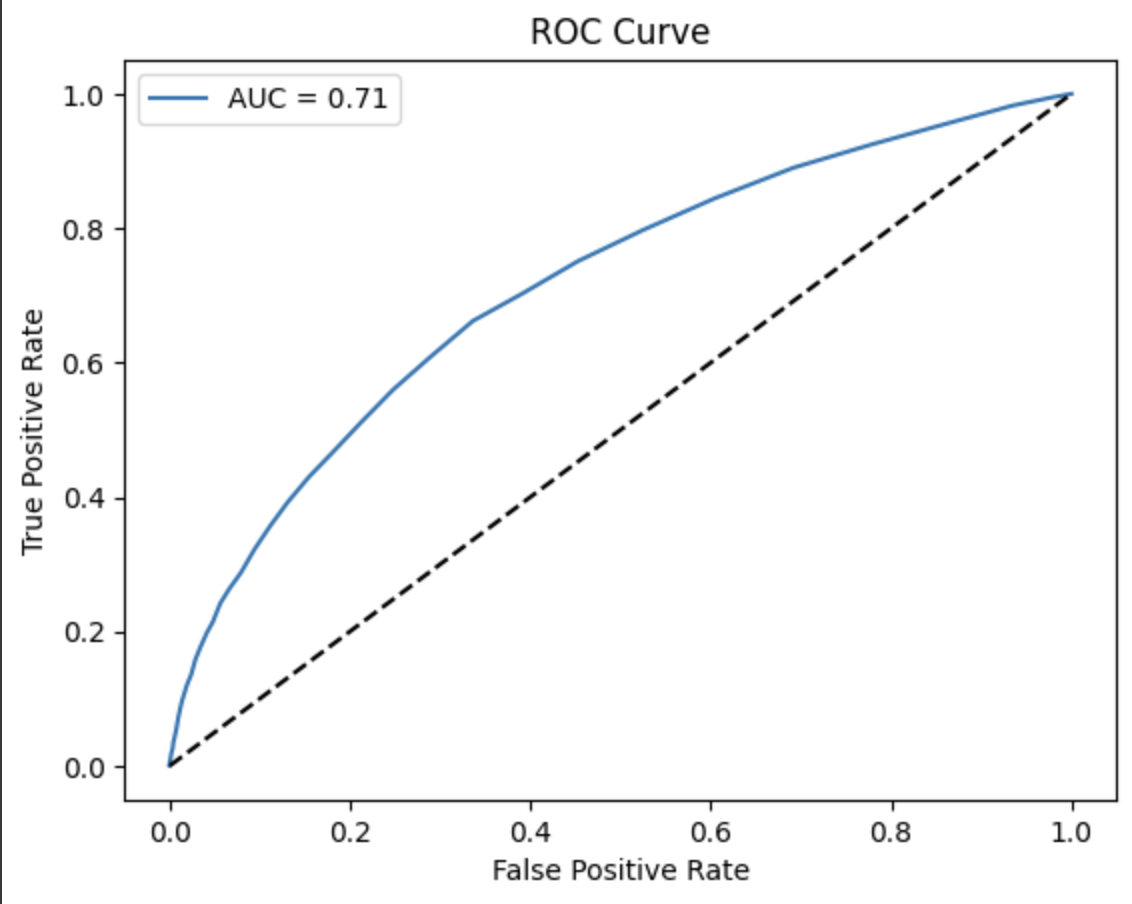
* 1. **Classification Report:**

****

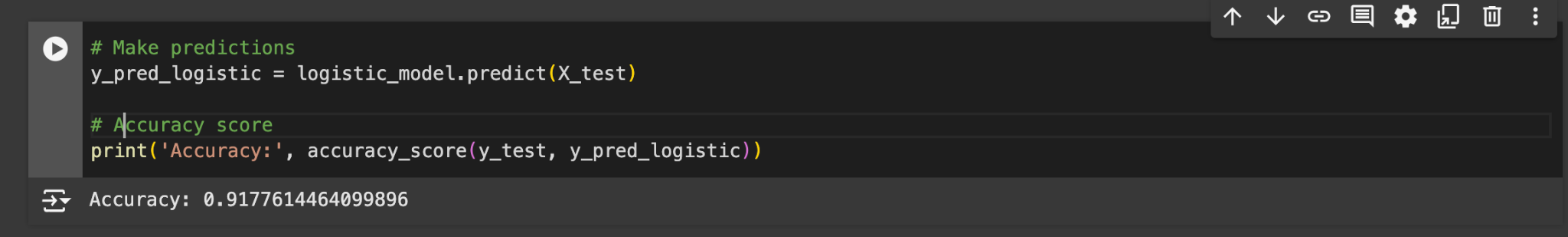
* 1. **Confusion Matrix:**

****

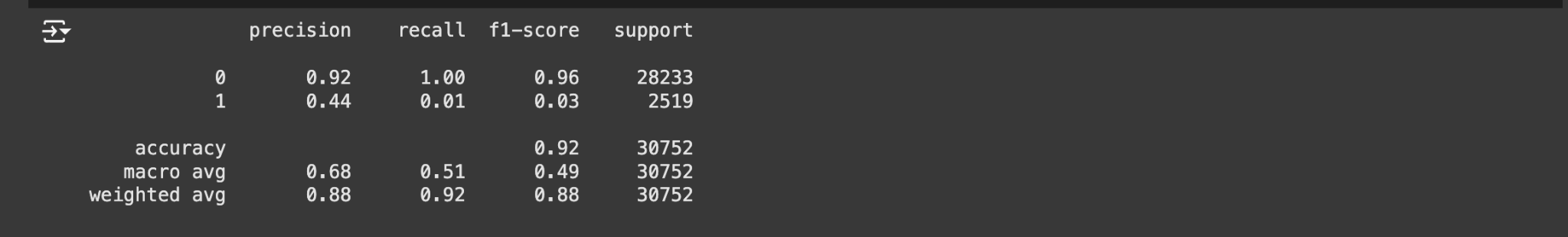
* 1. **ROC Curve:**

****

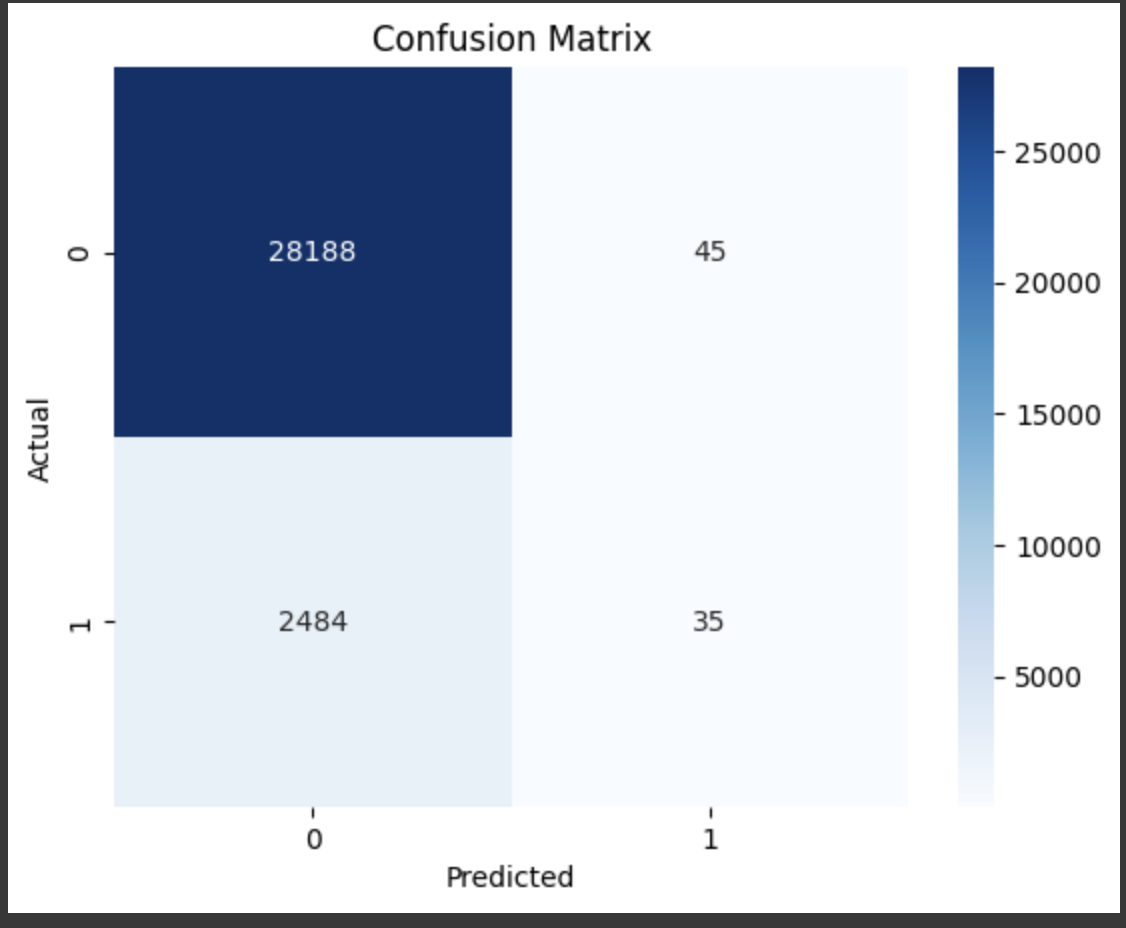
1. **Logistic Regression**
   1. **Accuracy:**

****

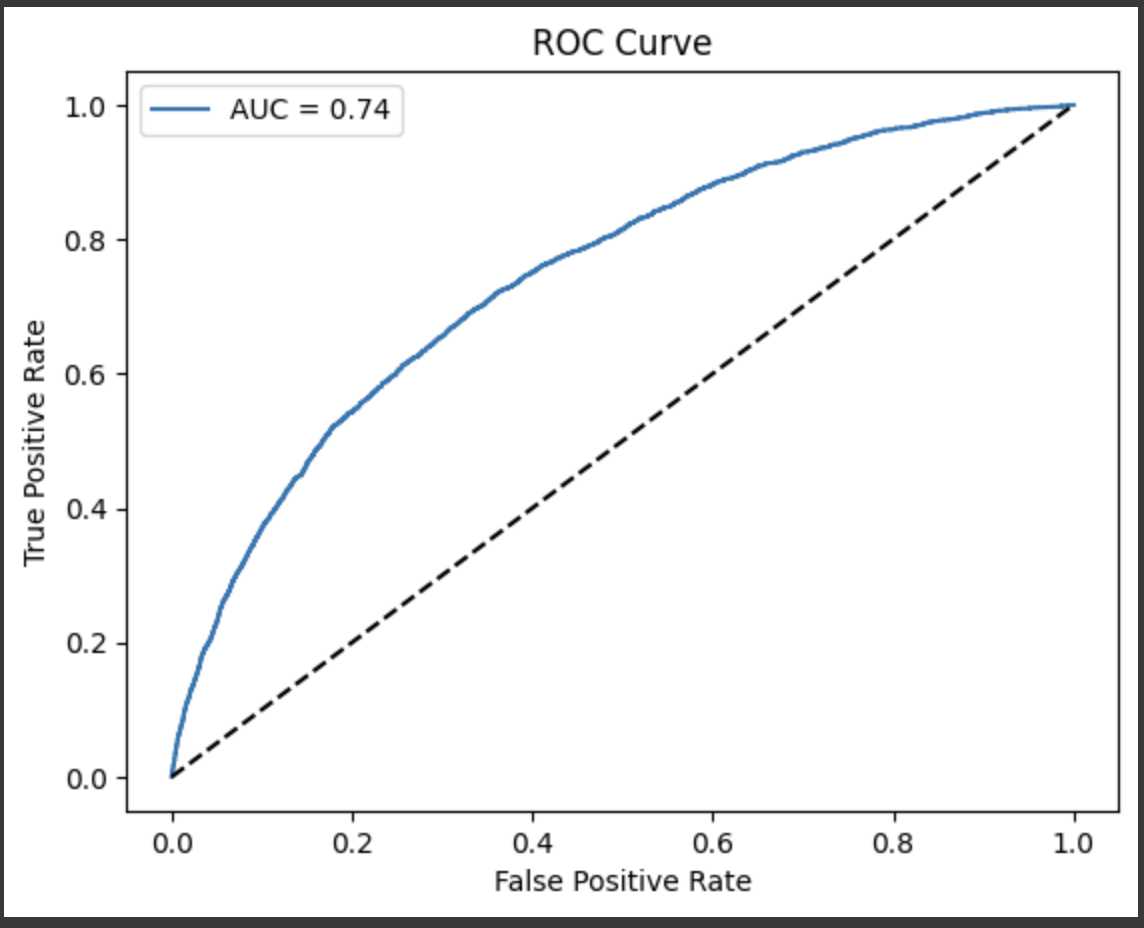
* 1. **Classification Report:**

****

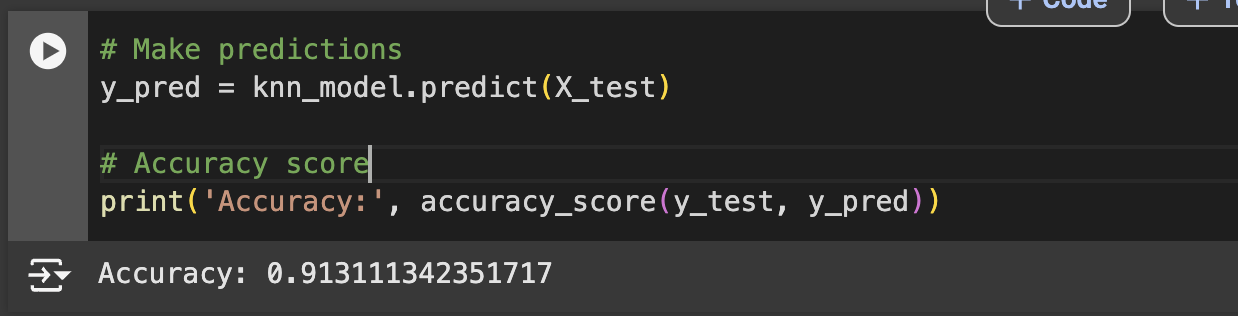
* 1. **Confusion Matrix:**

****

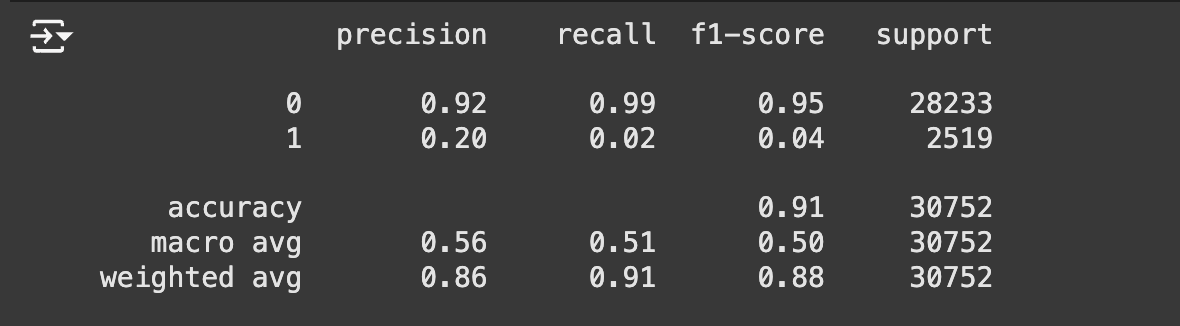
* 1. **ROC Curve:**

****

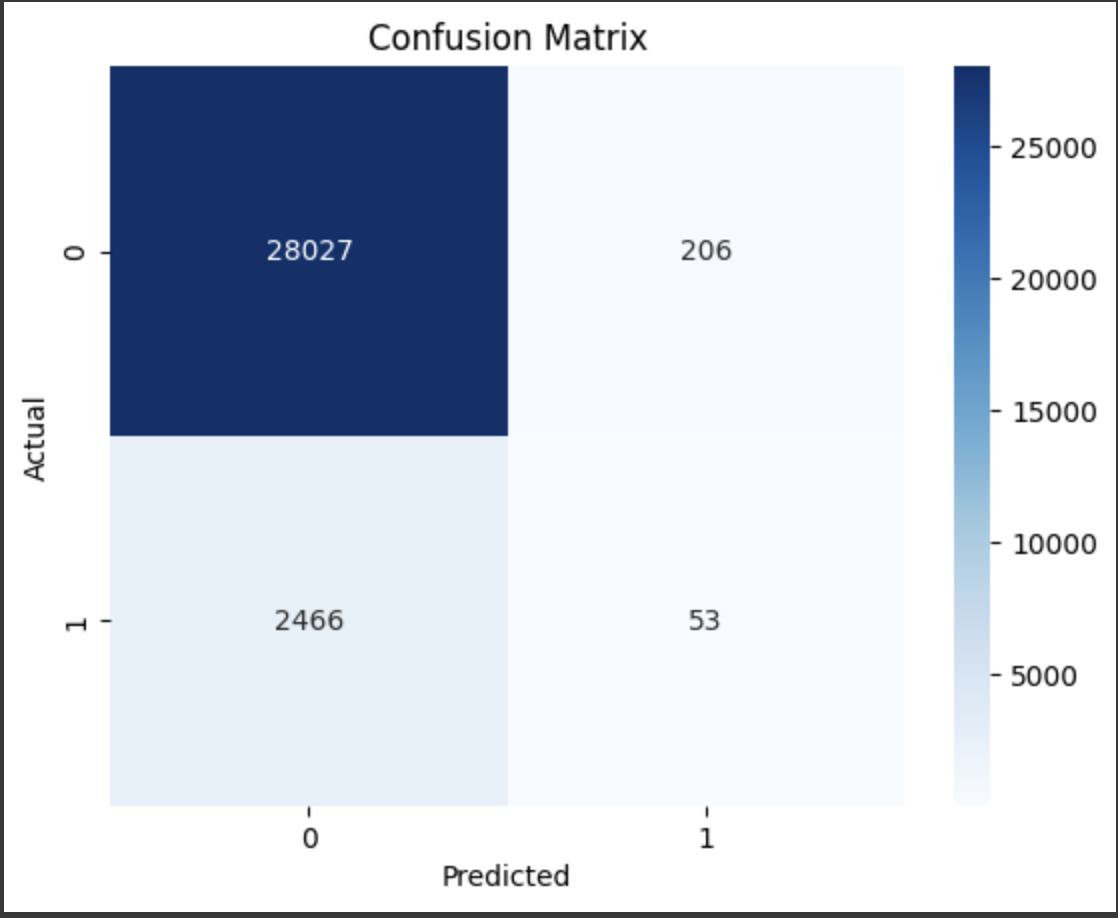
1. **KNN**
   1. **Accuracy Score:**

****

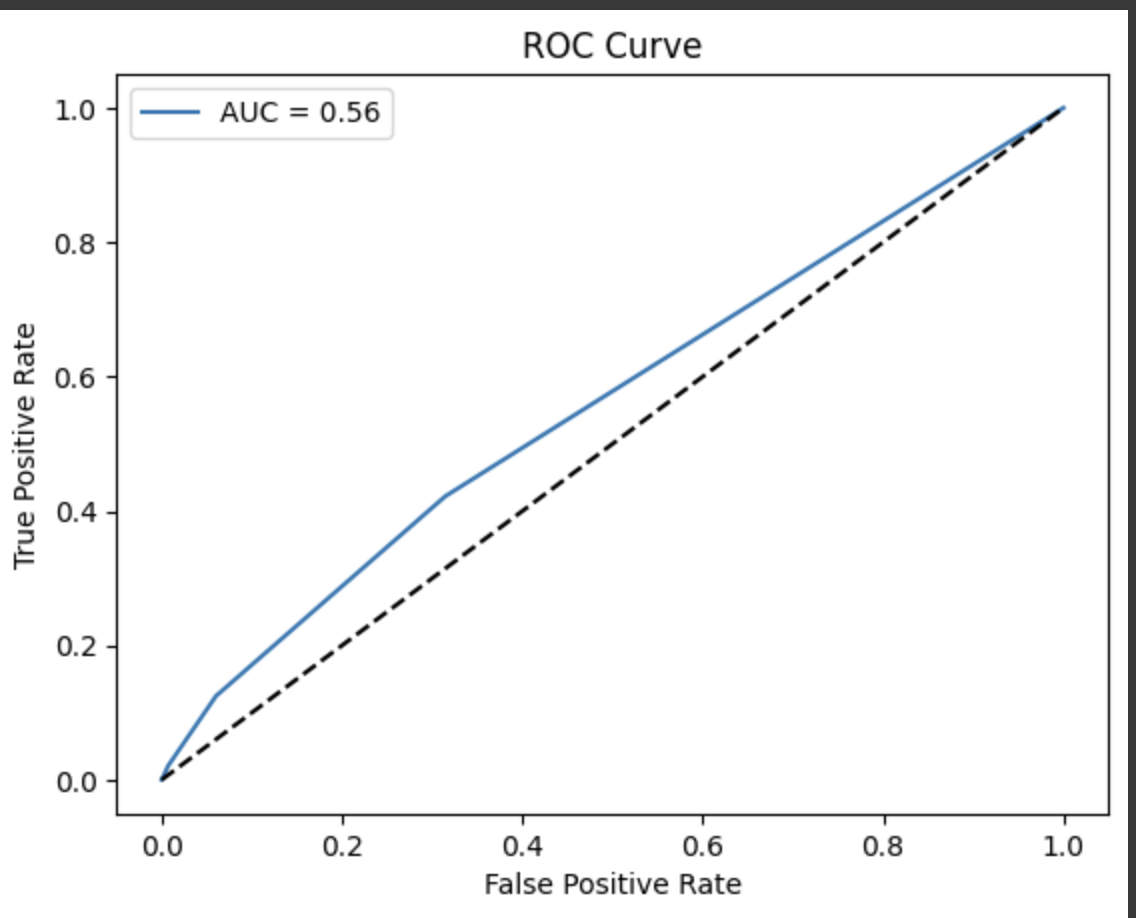
* 1. **Classification Report:**

****

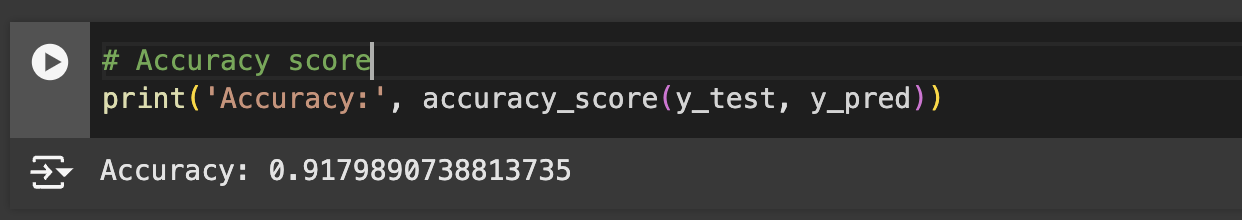
* 1. **Confusion Matrix:**

****

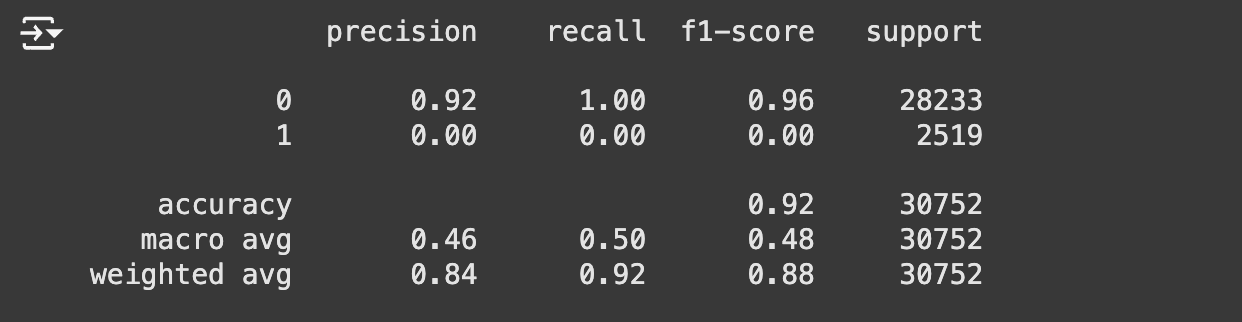
* 1. **ROC Curve:**

****

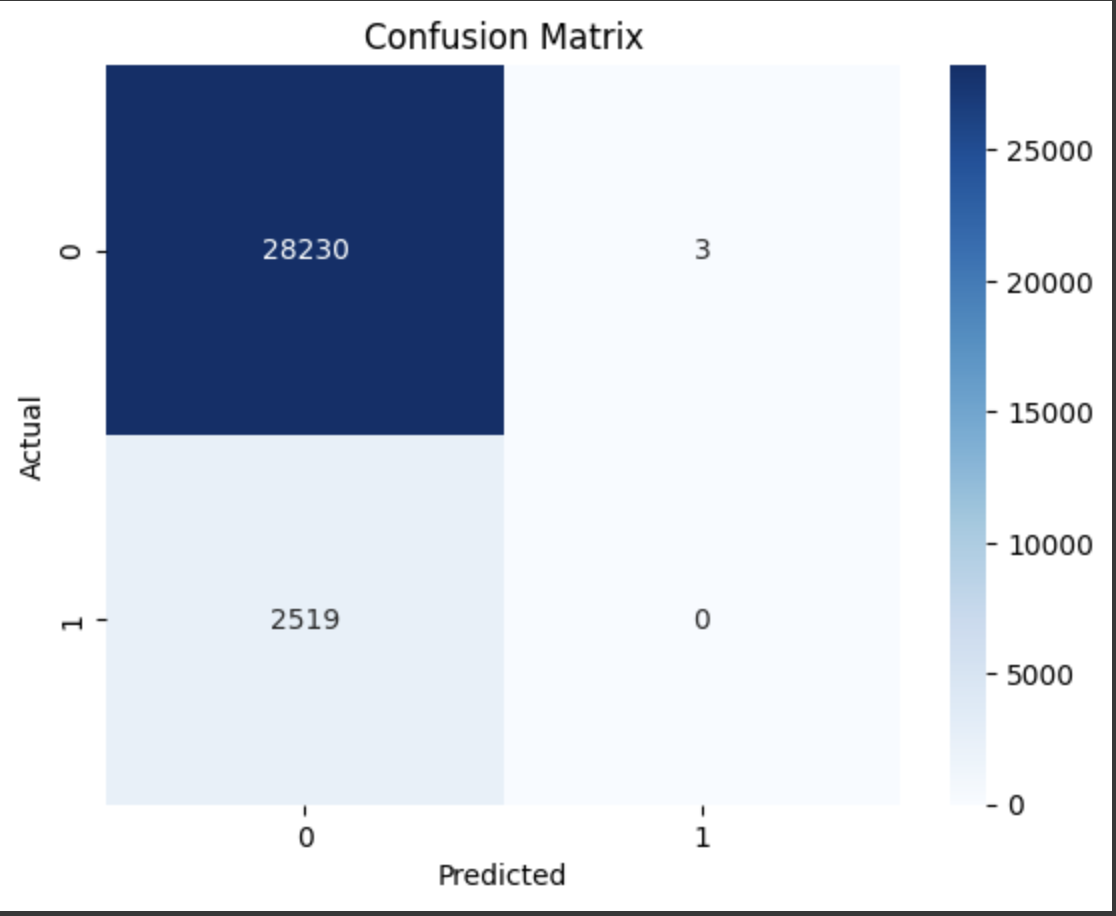
1. **Linear SVM:**
   1. **Accuracy Score:**

****

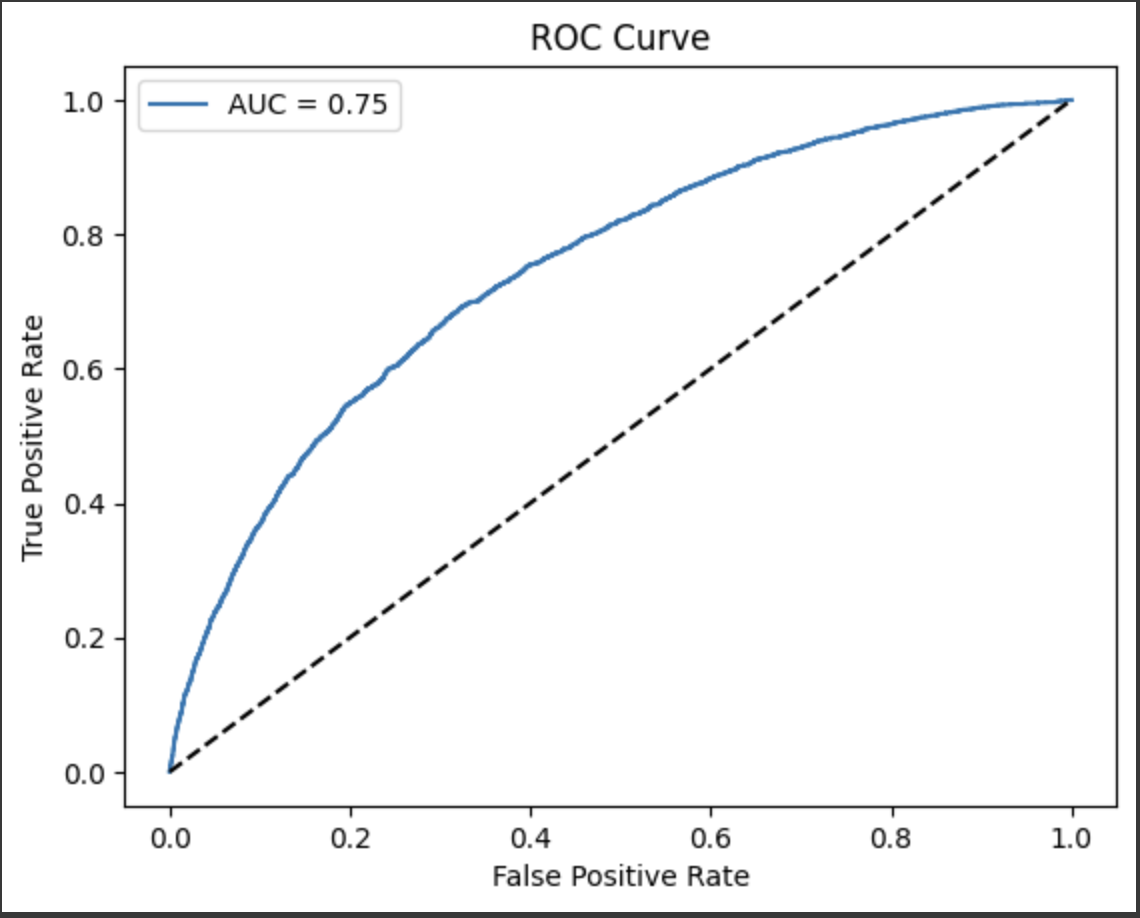
* 1. **Classification Report:**

****

* 1. **Confusion Matrix:**

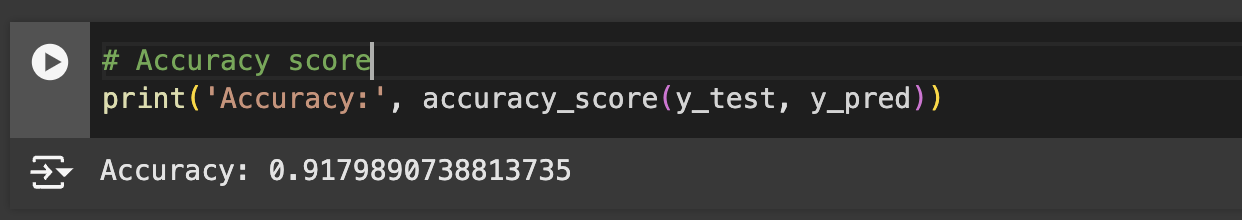
****

* 1. **ROC Curve:**

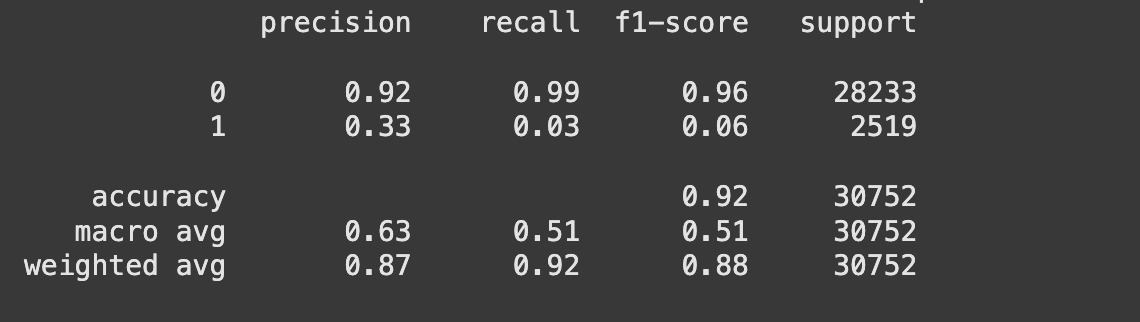
****

**5.Neural Networks:**

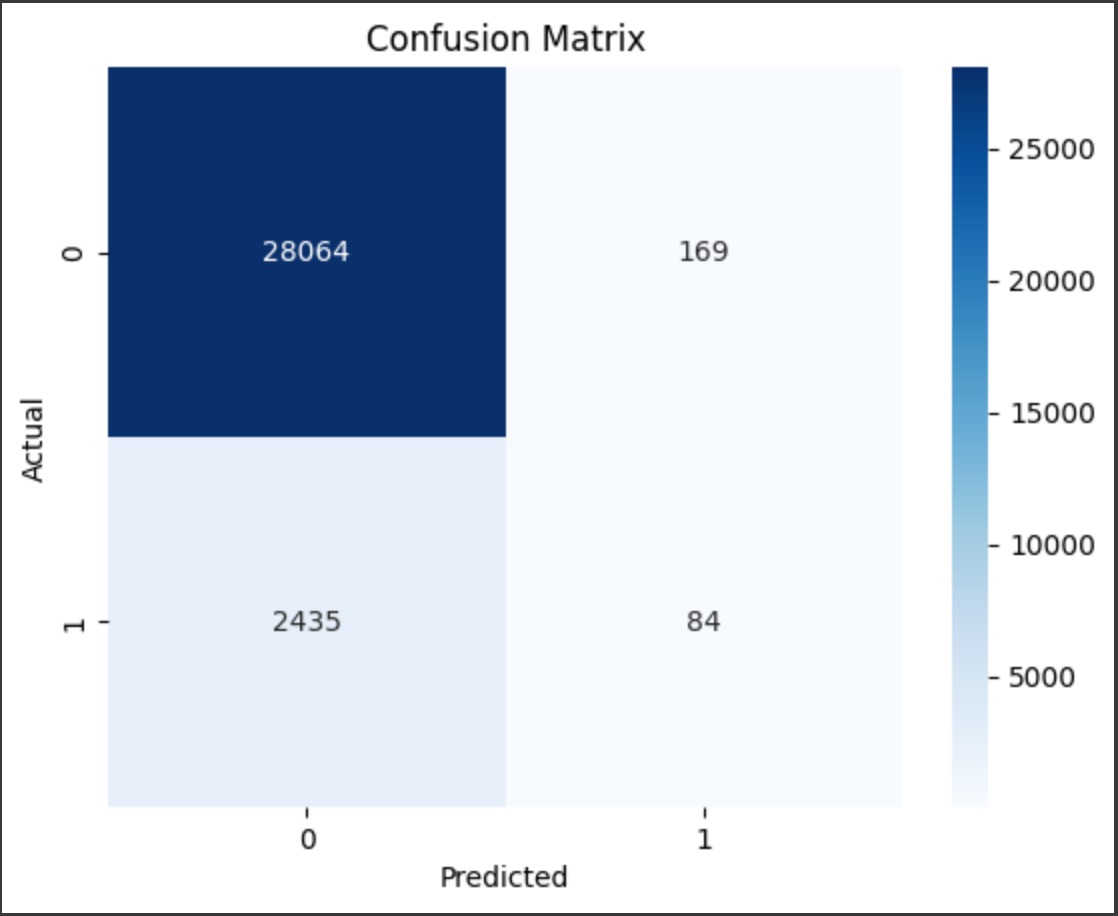
* 1. **Accuracy Score:**

****

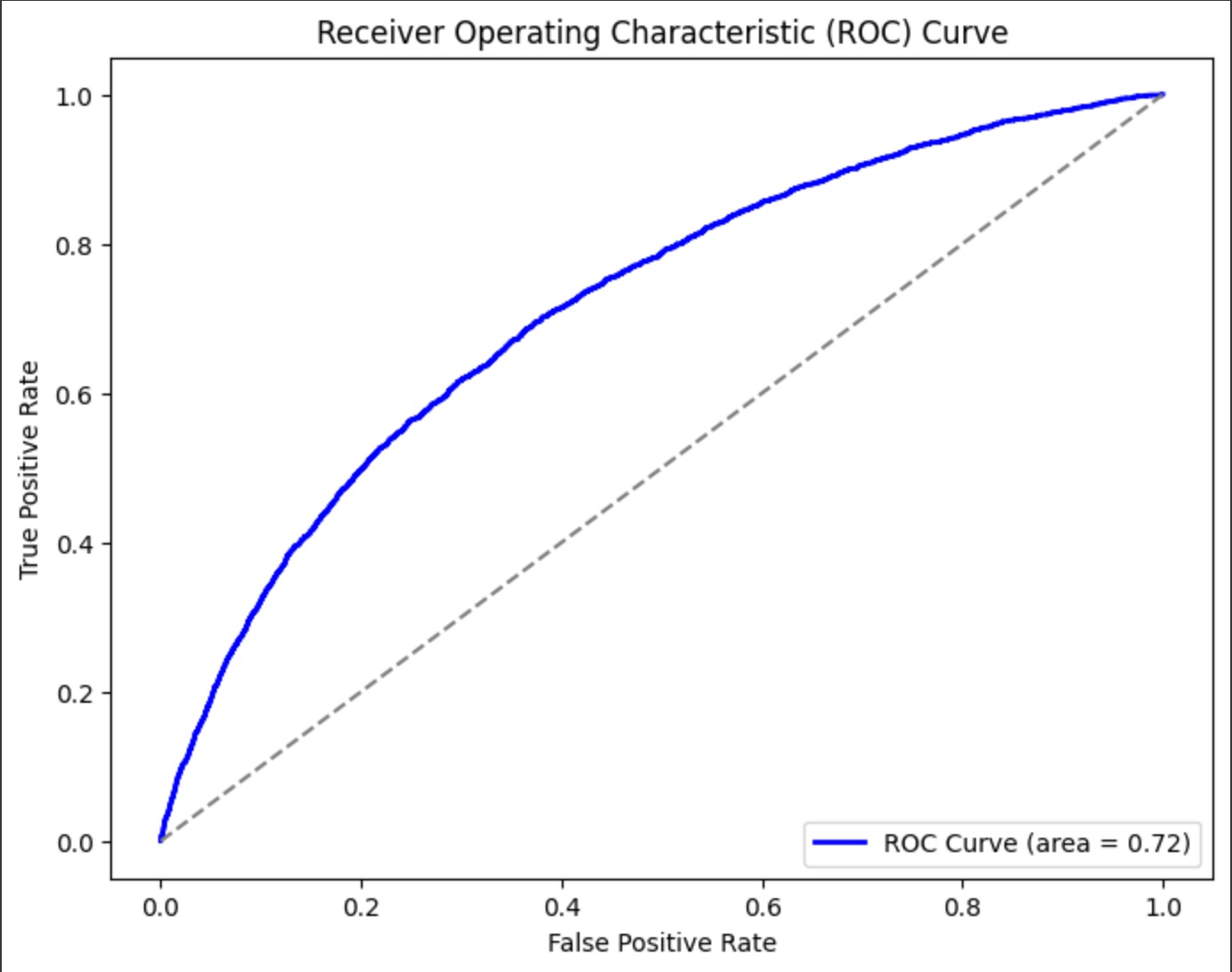
* 1. **Classification Report:**

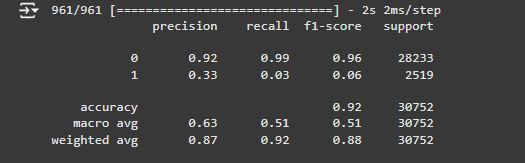
****

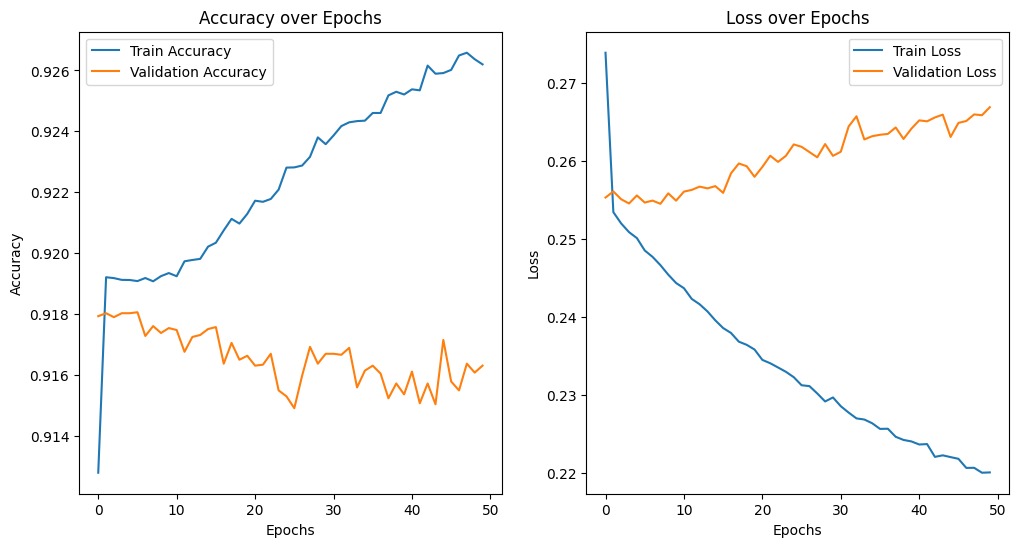
* 1. **Confusion Matrix:**

****

* 1. **ROC Curve:**

****

****

****

### **5.Feature Importance**

It was analysed to understand which variables had the greatest influence on predicting loan defaults. The top features included:

* **Debt-to-Income Ratio:** Indicates financial stress and is highly predictive of loan default.
* **Number of Previous Loans:** A higher number of past loans correlated with an increased likelihood of default.
* **Mobile Payment Behavior:** Telco data provided insight into applicants' non-traditional financial behaviour, which was also an important predictor.

**6.Challenges**

* **Class Imbalance**: The dataset likely suffers from an imbalance between default and non-default applicants, which could skew model performance, especially in terms of **Recall**. Models like **Random Forest** and **Logistic Regression** handled this imbalance better, but **KNN** and **Linear SVM** struggled.
* **Feature Selection and Engineering**: Key features like the **debt-to-income ratio** and **telco payment history** had significant impact on model performance. Proper feature engineering, including the creation of derived features, was essential in improving predictive accuracy.

**7.Conclusion**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Model** | **Accuracy** | **Precision** | **Recall** | **F1-Score** | **AUC-ROC** |
| **Random Forest** | 0.91815 | 0.90 | 0.92 | 0.88 | 0.71 |
| **Logistic Regression** | 0.91776 | 0.88 | 0.92 | 0.88 | 0.74 |
| **K-Nearest Neighbors** | 0.91311 | 0.86 | 0.91 | 0.88 | 0.56 |
| **Linear SVM** | 0.91798 | 0.84 | 0.92 | 0.88 | 0.75 |
| **Neural Networks** | 0.92 | 0.87 | 0.92 | 0.88 | 0.72 |

### 

* **Random Forest** was the best performer in terms of accuracy, precision, and recall, with **Logistic Regression** following closely behind in overall performance. Both models can be considered strong candidates for deployment, with **Random Forest** being more robust and flexible in handling complex relationships.
* **KNN** struggled with distinguishing defaults from non-defaults, as indicated by its low **AUC-ROC**. This highlights the challenges of using KNN for high-dimensional data like this, which may suffer from the "curse of dimensionality."
* **Linear SVM,Neural Networks** showed competitive **Recall**, but its lower **AUC-ROC** indicates it was less effective at distinguishing between the two classes. **Linear SVM** may benefit from more tuning or the application of non-linear kernels for more complex decision boundaries.
* **Logistic Regression** satisfies and outperforms Random Forest ,Neural Network in terms of AUC while recall for 3 of them are same

**Team:**

* Srinath Bulusu-1002197525
* Pranay Mohan Kanakabandi -1002198613
* Manas Pavan Sai Kallaganti - 1002209866
* Sai Prabhakar Gattu - 1002209363