Milestone 1 Report

Task 1: Study the characteristics of the data and identify the quality issues in the selected.

Data characteristics

- This dataset focuses on the classification of who can earn money more than 50K per year.
- This dataset is an imbalanced class, class 0 contains 24720 values and class 1 contains 7841 values.
- Some categorical features have missing values and have many types in categorical column.
- Many features can be used in feature engineering to make a new feature.

Quality issues in the dataset

- Bias: Imbalance class effective to the model when training
- Fairness and reliability: Because the model is bias, and unfairness makes the model unconvincing.
- Performance: Bad quality of the dataset effect to the model performance.

Task 2: Define the goals and a suitable measure for the quality issues.

- Improve overall and specify metric performance

Measures:

- F1-Score and Precision, because this model needs to classify who can earn more than 50K per year then it will focus on class 1.

Progress:

- More EDA in the dataset.
- Apply sample techniques to handle imbalanced class.
- Create models using new preparation data and compare them with baseline models.

Task 3: Explore different kinds of machine learning models developed with different modeling techniques. Then, choose the machine learning techniques, implement the models using scikit-learn, and train the models.

Working step:

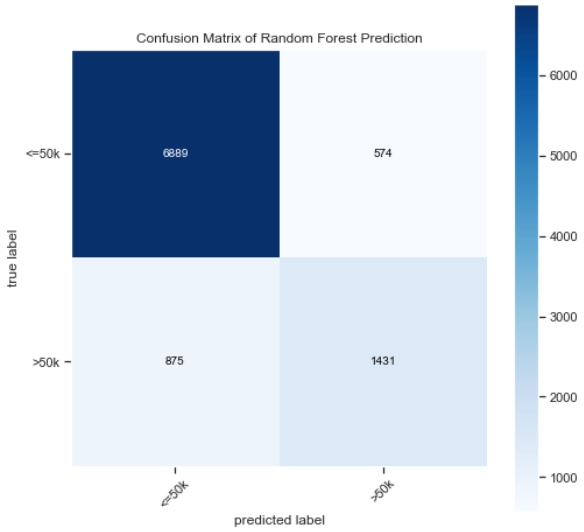
- EDA the dataset
- Prepare data
- Train and test data
- Compute model performance

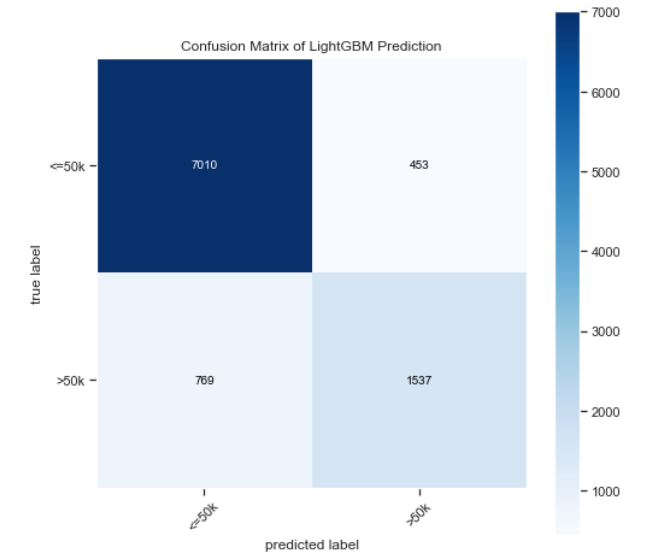
For this part, we have studied different 6 machine learning models for training and testing the dataset and each model prepares data the same method. The performance of baseline models and confusion matrix.

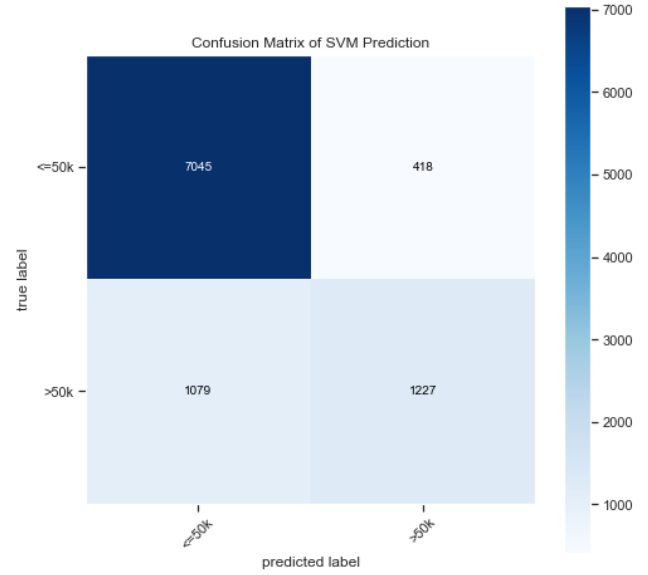
Table 1: Baseline model

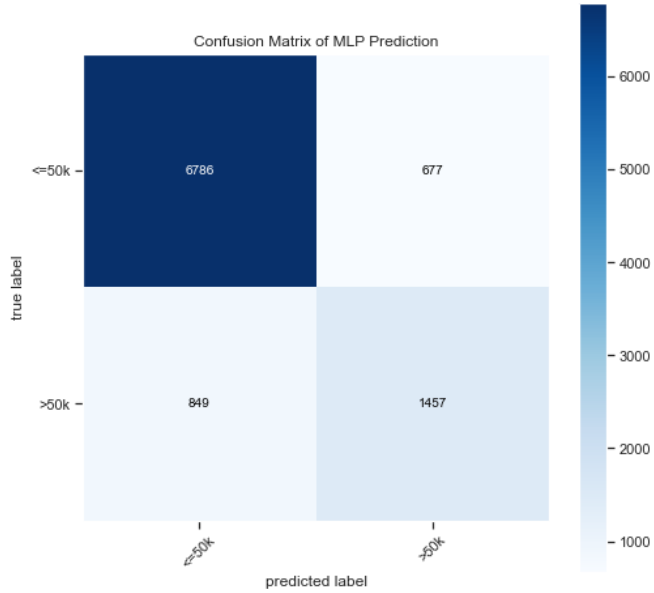
Model	Accuracy (%)	Recall (%)	Precision (%)	F1-Score (%)	Implementer
Random Forest	85.1674	85.1674	84.6325	84.7962	Rew
LightGBM	87.4910	87.4910	87.0745	87.1607	Rew
SVM	84.6760	84.6760	83.8554	83.7191	B
MLP	84.3792	84.3792	84.0164	84.1656	B
Logistic Regression	82.3012	82.3012	81.0242	80.7996	Joey
K-NN	83.1201	83.1201	82.6049	82.8066	Joey

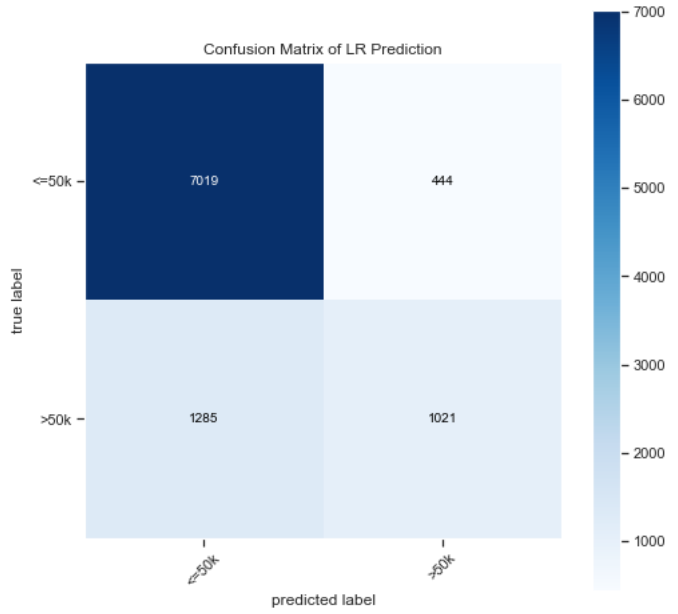
Here is confusion matric from these models in table 1.

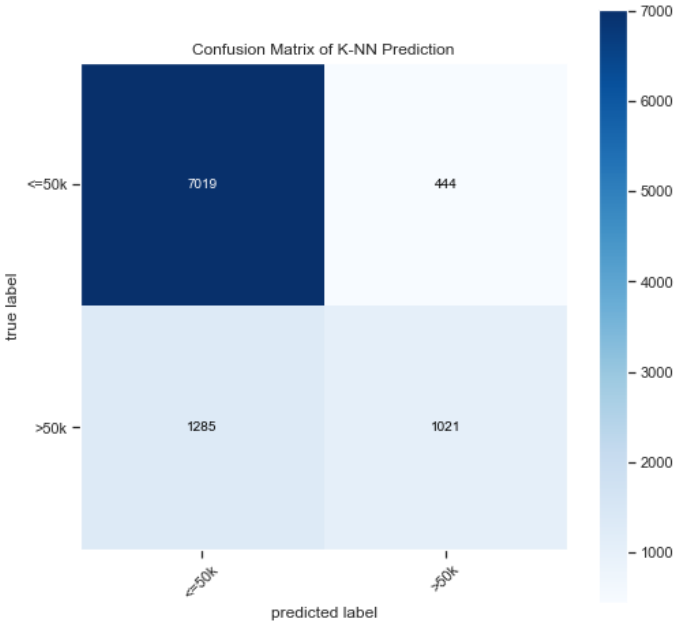



Confusion Matrix of LightGBM Prediction
<=50k
>50k
true label
7010
453
769
1537
<=50k
>50k
predicted label
7000
6000
5000
4000
3000
2000
1000


Confusion Matrix of SVM Prediction
<=50k
>50k
true label
7045
418
1079
1227
<=50k
>50k
predicted label
7000
6000
5000
4000
3000
2000
1000


Confusion Matrix of MLP Prediction
<=50k
>50k
true label
6786
677
849
1457
<=50k
>50k
predicted label
6000
5000
4000
3000
2000
1000


Confusion Matrix of LR Prediction
<=50k
>50k
true label
7019
444
1285
1021
<=50k
>50k
predicted label
7000
6000
5000
4000
3000
2000
1000


Confusion Matrix of K-NN Prediction
<=50k
>50k
true label
7019
444
1285
1021
<=50k
>50k
predicted label
7000
6000
5000
4000
3000
2000
1000