Task 3 – Feature Engineering for Model Gains

1.Title Page Title: Feature Engineering for Model Gains

Course: FutureXcel – Machine Learning

Week: 3 Prepared by: Mahnoor khan

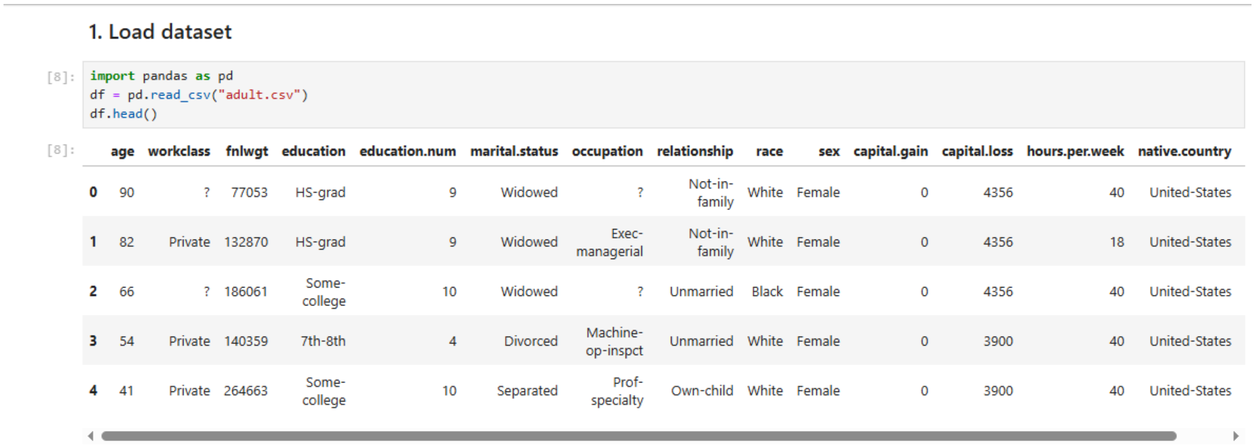
2. Introduction

The objective of this task is to apply feature engineering on the Adult Census Income dataset to improve the predictive performance of a machine learning model.

Baseline model: Logistic Regression using original numeric features.

Improved model: Random Forest Classifier using engineered features.

3. Dataset Overview



1. Load dataset

```
import pandas as pd
df = pd.read_csv("adult.csv")
df.head()
```

	age	workclass	fnlwgt	education	education.num	marital.status	occupation	relationship	race	sex	capital.gain	capital.loss	hours.per.week	native.country
0	90	?	77053	HS-grad	9	Widowed	?	Not-in-family	White	Female	0	4356	40	United-States
1	82	Private	132870	HS-grad	9	Widowed	Exec-managerial	Not-in-family	White	Female	0	4356	18	United-States
2	66	?	186061	Some-college	10	Widowed	?	Unmarried	Black	Female	0	4356	40	United-States
3	54	Private	140359	7th-8th	4	Divorced	Machine-op-inspct	Unmarried	White	Female	0	3900	40	United-States
4	41	Private	264663	Some-college	10	Separated	Prof-specialty	Own-child	White	Female	0	3900	40	United-States

. Baseline Model (Before Feature Engineering)

features used: age, fnlwgt, education.num, capital.gain, capital.loss, hours.per.week ⍰ Model: Logistic Regression ⍰

Train/test split: 80/20 ===Baseline Accuracy (Before FE) === Accuracy: 0.8188238906801781

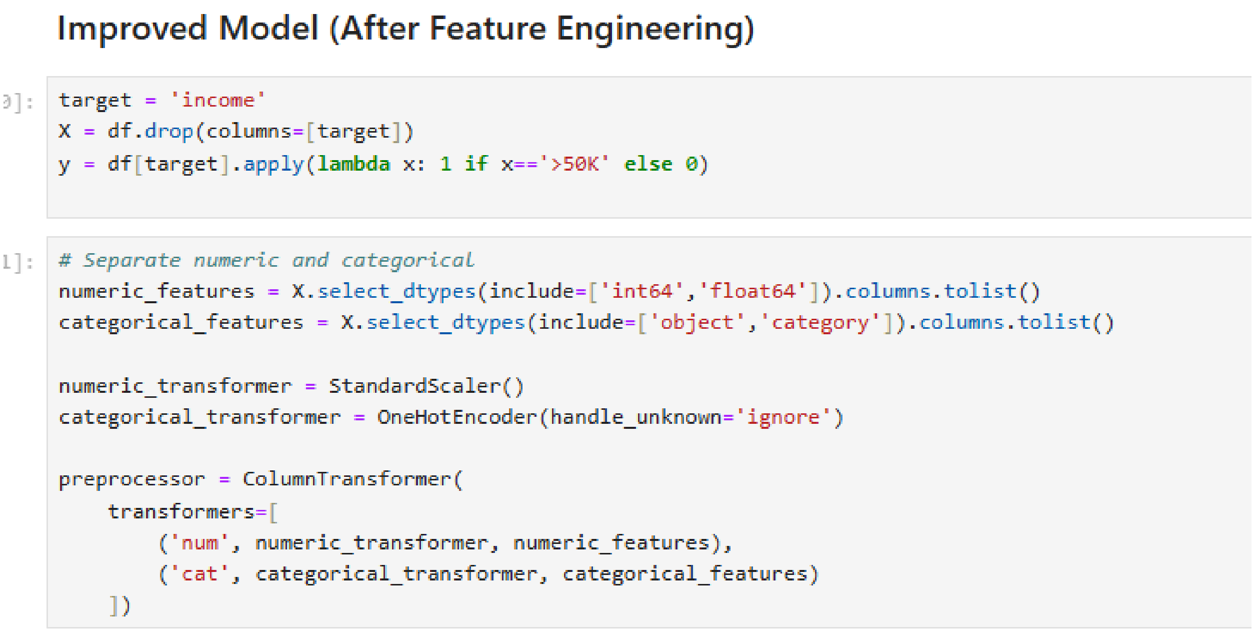
5. Feature Engineering (10 Features)

6. Improved Model (After Feature Engineering)

Features used: Original features + 10 engineered features

Model: Random Forest Classifier

Train/test split: 80/20 === Improved Accuracy (After FE) === Accuracy: 0.855366190695532



Improved Model (After Feature Engineering)

```
target = 'income'
X = df.drop(columns=[target])
y = df[target].apply(lambda x: 1 if x=='>50K' else 0)
```

```
# Separate numeric and categorical
numeric_features = X.select_dtypes(include=['int64','float64']).columns.tolist()
categorical_features = X.select_dtypes(include=['object','category']).columns.tolist()

numeric_transformer = StandardScaler()
categorical_transformer = OneHotEncoder(handle_unknown='ignore')

preprocessor = ColumnTransformer(
    transformers=[
        ('num', numeric_transformer, numeric_features),
        ('cat', categorical_transformer, categorical_features)
    ])
```

7. Comparison (Before vs After Feature Engineering) Accuracy Comparison:

Model Baseline (Before FE)

Improved (After FE) Accuracy 0.8188238906801781 0.855366190695532

8. Conclusion

Feature engineering improved the predictive performance of the model. ⍰ Key engineered features such as age_squared, net_capital, and age_hours contributed to better accuracy. Random Forest effectively handled non-linear relationships and categorical variables.