Implementation of Multinomial Logistic Regression to Predict Credit Grade In Irish-Loan-Dataset

Azis Muslim
Data Scientist

My Professional Background

Data Infrastructure Engineer at Moladin
From Dec 1st 2022 until Feb 7th 2023

Key Responsibilities:

- Maintaining Data Infrastructure
- Building Data Infrastructure in Kubernetes
- Coordinating with Data Engineer for Data Application Research
- Granting Database Access for those who were needed by jira tickets
- Building cron-job to clean up kubernetes from failed pods and not running pods
- Debugging Airflow to tracks jobs error
- Building Gitlab CI/CD for job deployment in Apache Airflow
- Setting up GCP with Terraform

Research Steps

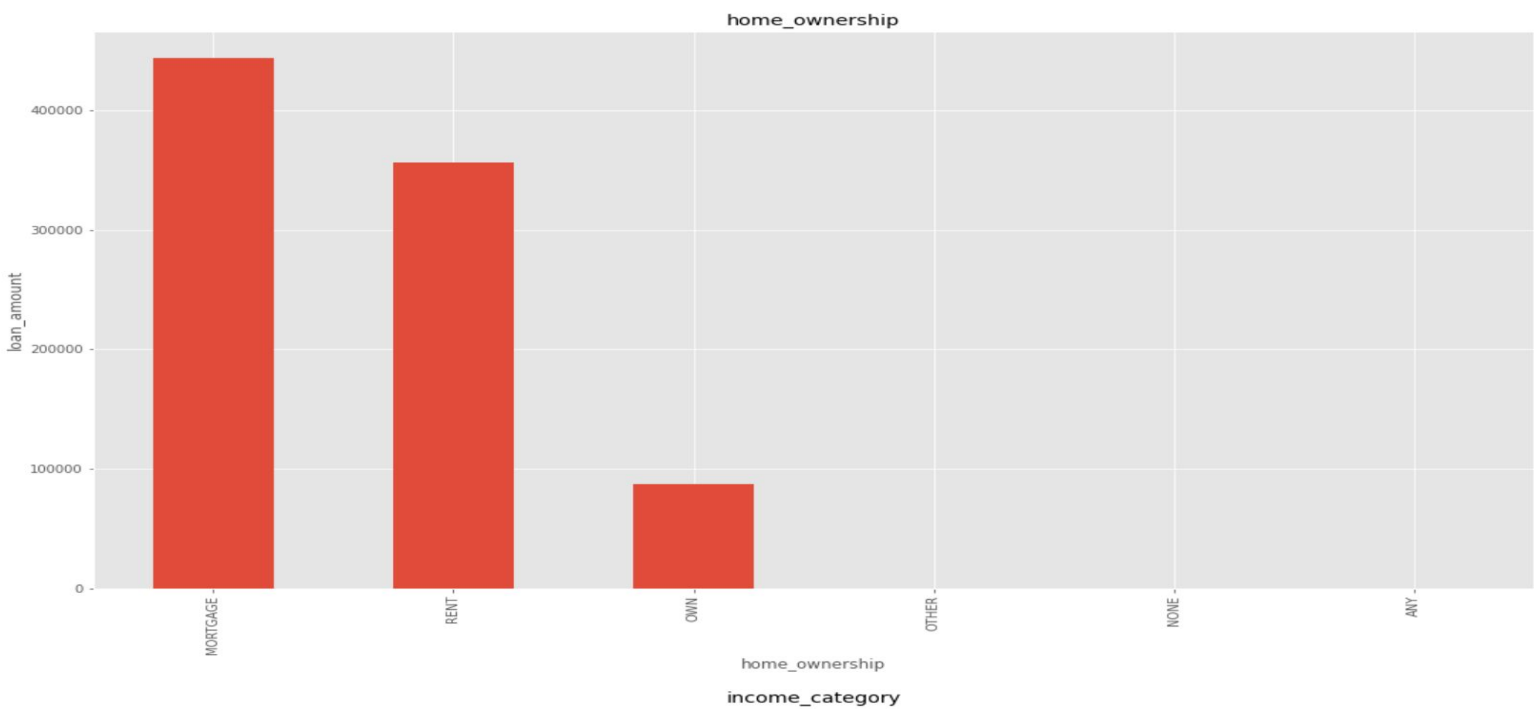
1. Preparing the data and process the data from dirty to clean
2. Identify the characteristics of datasets and deciding of which type of machine learning algorithm that would be implemented for classification
3. Implementing Principal component analysis for high dimensional features dataset
4. Grasp the concept of logistic regression for classification to interpret the result of machine learning data processing
5. Evaluating the machine learning model

Dataset Overview

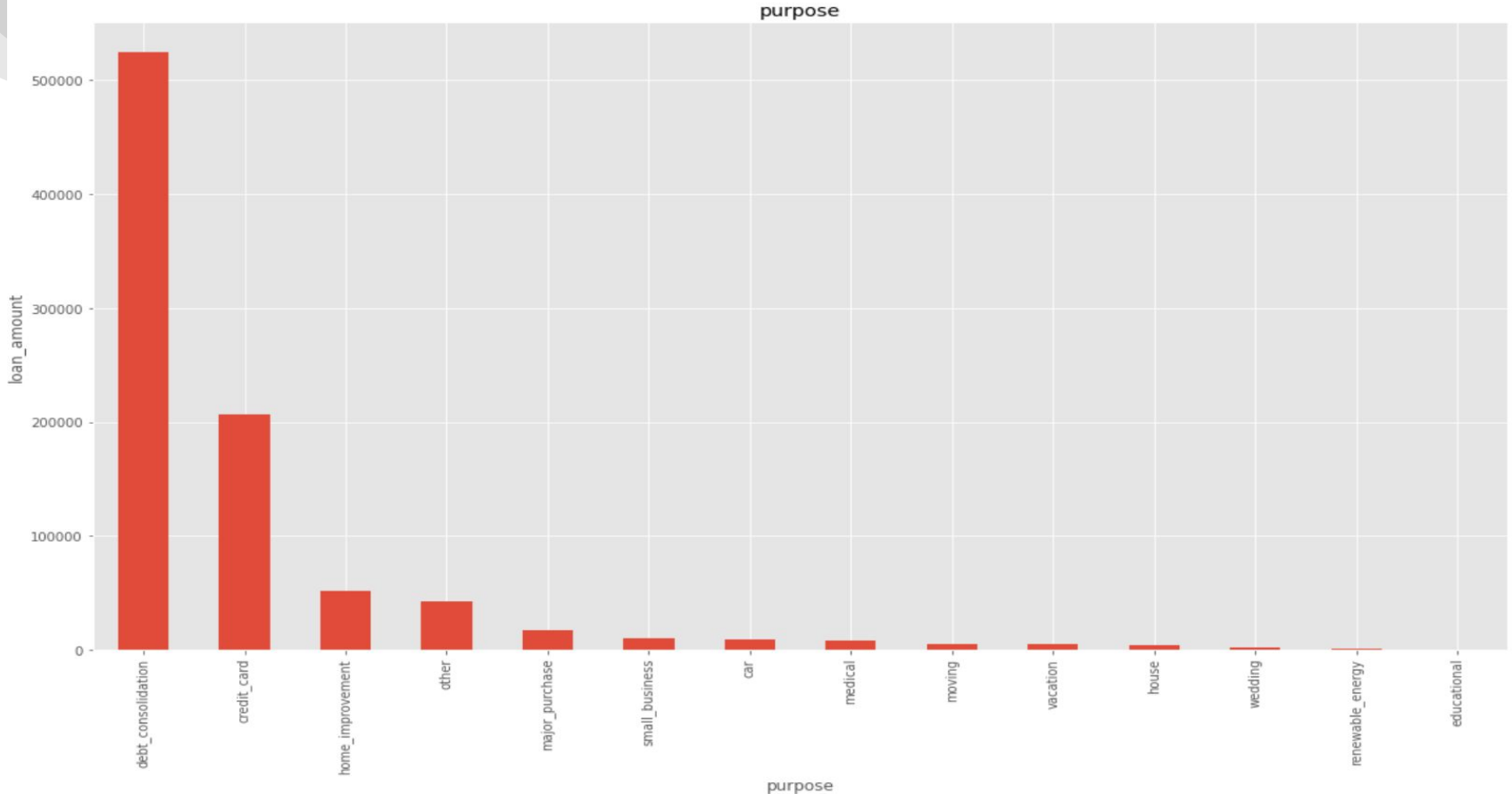
	id	year	issue_d	final_d	emp_length_int	home_ownership	home_ownership_cat	income_category	a
0	1077501	2011	01/12/2011	1012015	10.0	RENT	1	Low	
1	1077430	2011	01/12/2011	1042013	0.5	RENT	1	Low	
2	1077175	2011	01/12/2011	1062014	10.0	RENT	1	Low	
3	1076863	2011	01/12/2011	1012015	10.0	RENT	1	Low	
4	1075358	2011	01/12/2011	1012016	1.0	RENT	1	Low	
...	...	...	...	...	...	...	...	...	
887374	36371250	2015	01/01/2015	1012016	8.0	RENT	1	Low	
887375	36441262	2015	01/01/2015	1012016	10.0	MORTGAGE	3	Low	
887376	36271333	2015	01/01/2015	1012016	5.0	RENT	1	Low	
887377	36490806	2015	01/01/2015	1012016	1.0	RENT	1	Low	
887378	36271262	2015	01/01/2015	1012016	10.0	RENT	1	Low	

887379 rows × 30 columns

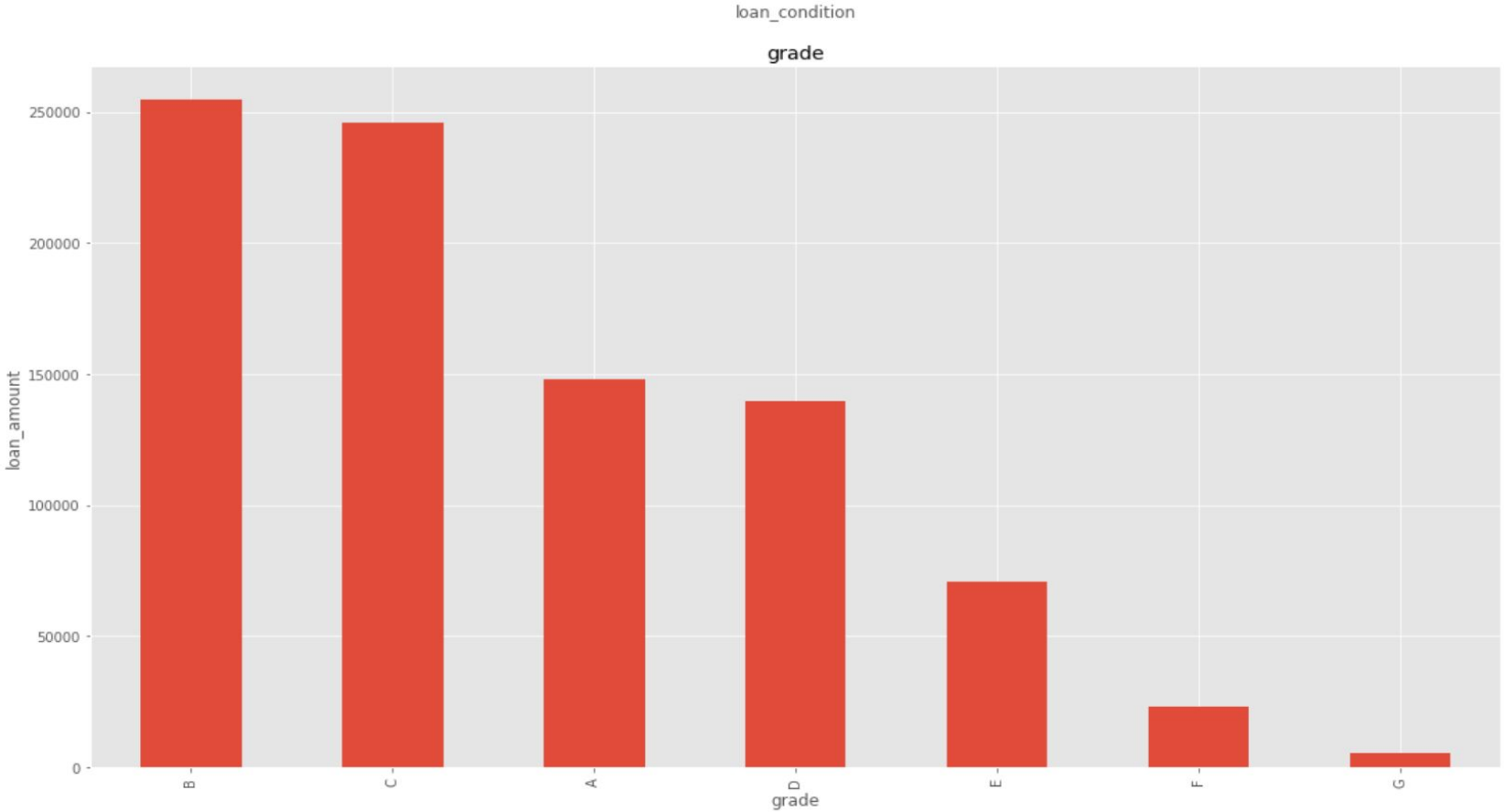
Exploratory Data Analysis



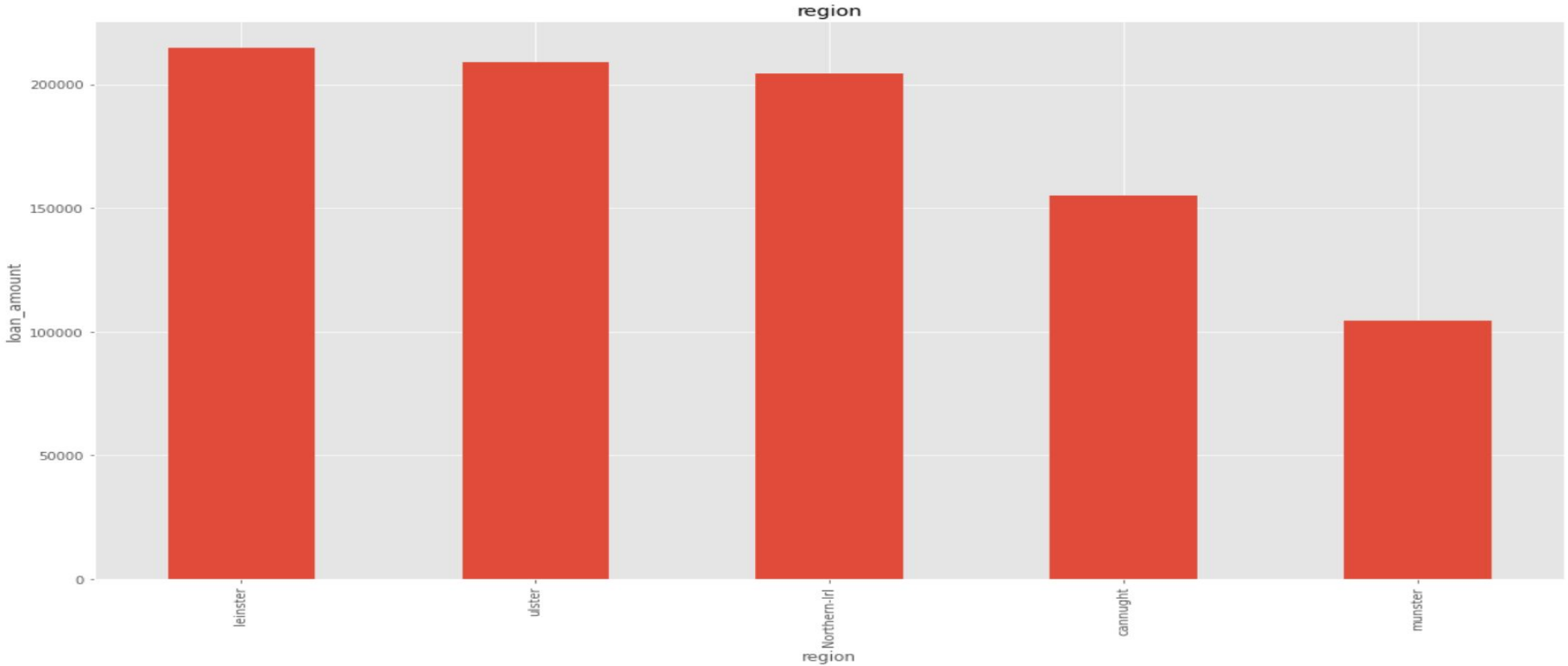
Debt Purpose



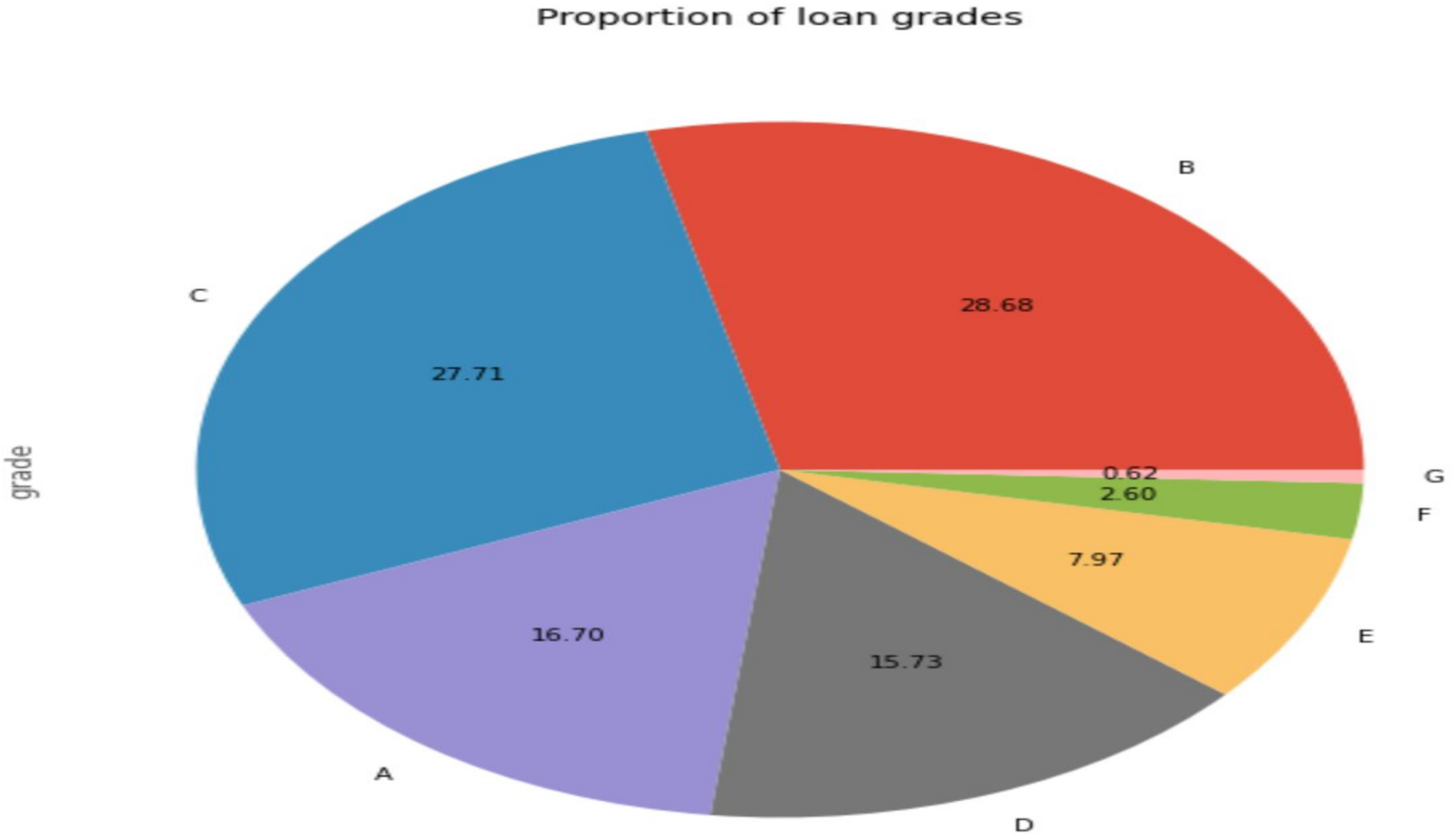
Loan Amount Based on Grades



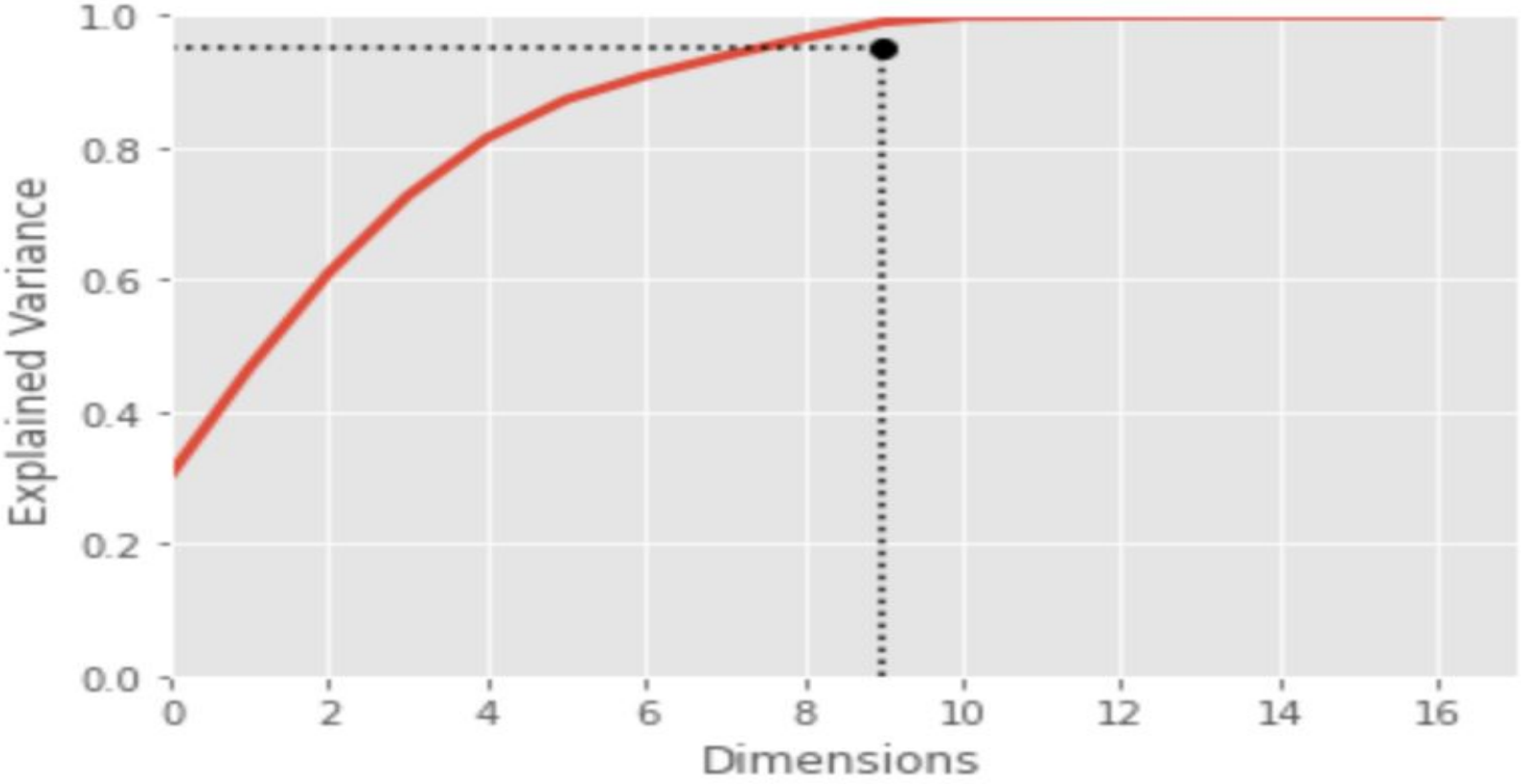
Loan Amount Based on Region



Proportion of Loan Grades



Dimensionality Reduction with PCA



Defining Pipeline for Logistic Regression

```
# Defining the pipeline

pipe_lr = Pipeline([
    ('min_max_scaler_lr', MinMaxScaler()),
    ('PCA_lr', PCA(n_components=10)),
    ('model_lr', LogisticRegression(multi_class= 'multinomial'))
])

# creating list which containes pipeline

pipelines = [pipe_lr]
```

Classification Report

	precision	recall	f1-score	support
1	0.89	0.96	0.93	29791
2	0.75	0.89	0.82	50616
3	0.81	0.68	0.74	49270
4	0.74	0.76	0.75	27891
5	0.61	0.63	0.62	14115
6	0.40	0.07	0.12	4690
7	0.00	0.00	0.00	1101
accuracy			0.78	177474
macro avg	0.60	0.57	0.57	177474
weighted avg	0.76	0.78	0.76	177474

Conclusion

The machine learning result by using multinomial logistic regression was showed that it has good precision for label A, B, C, D which have precision score 0.89, 0.75, 0.81, and 0.74 respectively. In contrast the logistic regression model tend to have bad precision for the label E, F, G which have precision score at 0.61, 0.40, and 0.00. This happened because the was huge gap of imbalance data between each label in grade column as a target feature.

then we were doing data preprocessing to see the dataset characteristics and here were our findings:

- Proportion of bad loans were relatively small
- High income had the highest capability of payment followed by middle income and low income
- The highest total payment were at grade G, F, E, D
- Region with the lowest amount of loan was located at munster
- Region with the highest amount of loan was located at leinster
- High income borrower tend to have smaller amount of loan