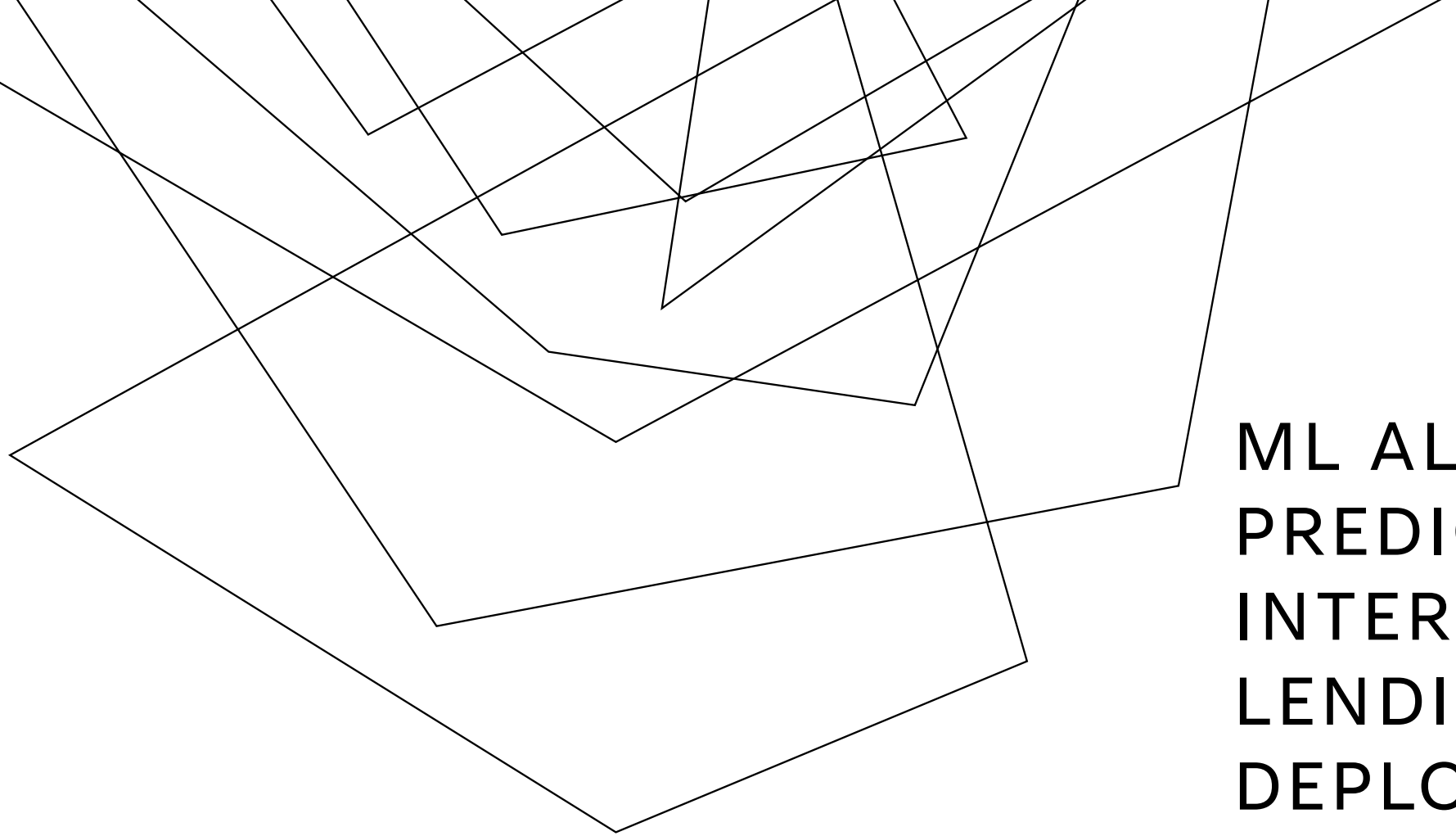


ML ALGORITHM PREDICITON OF INTEREST RATES FOR LENDING CLUB AND DEPLOYING IN THE AZURE CLOUD

Trishala Basti

Data Scientist

AGENDA

Introduction

Primary goals

Exploratory Data Analysis

Data cleaning

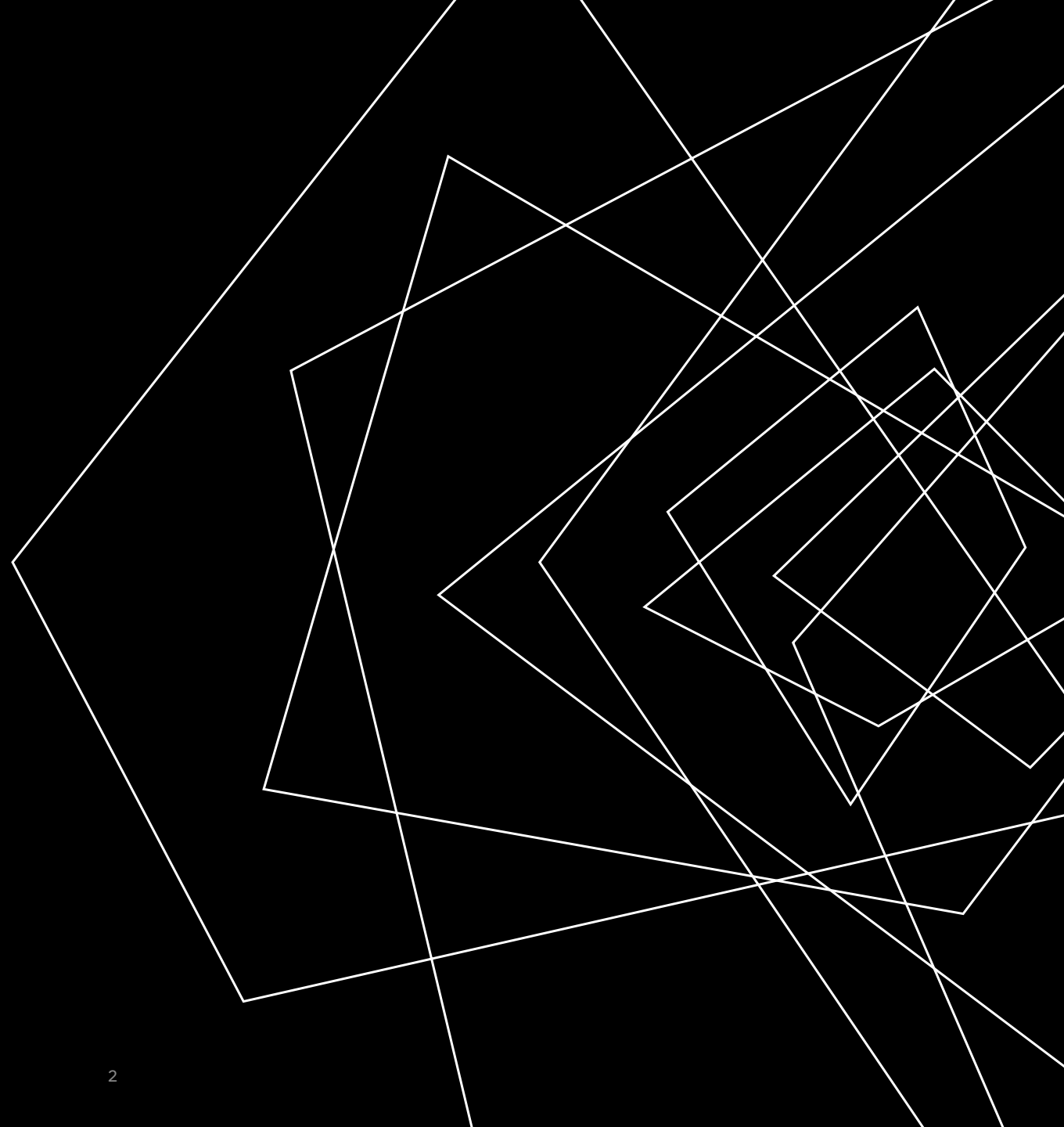
Feature Engineering

MODEL building with Different Algorithms

Model Scores

Evaluation

Final Inference



BUSINESS UNDERSTANDING AND PROJECT GOAL

Buisness Understanding:

Lending Club: the company wants to offer potential customers an online tool to predict potential interest rates based on the purpose and other variables

GOAL: use the dataset provided for different users with different properties and predict the interest rate for the new data.

EXPLORATORY DATA ANALYSIS

VARIABLES:

We took a look at the different variables required to predict or which determines the target variable. Each of them has been analysed to find the null values and outliers.

int_rate: This is the Target variable which is the Interest Rate on the loan

loan_amnt: **F**eature which is The listed amount of the loan applied for by the borrower

term : The number of payments on the loan. Values are in months and can be either 36 or 60

Grade: Assigned loan grade based on the income and other stats of the user

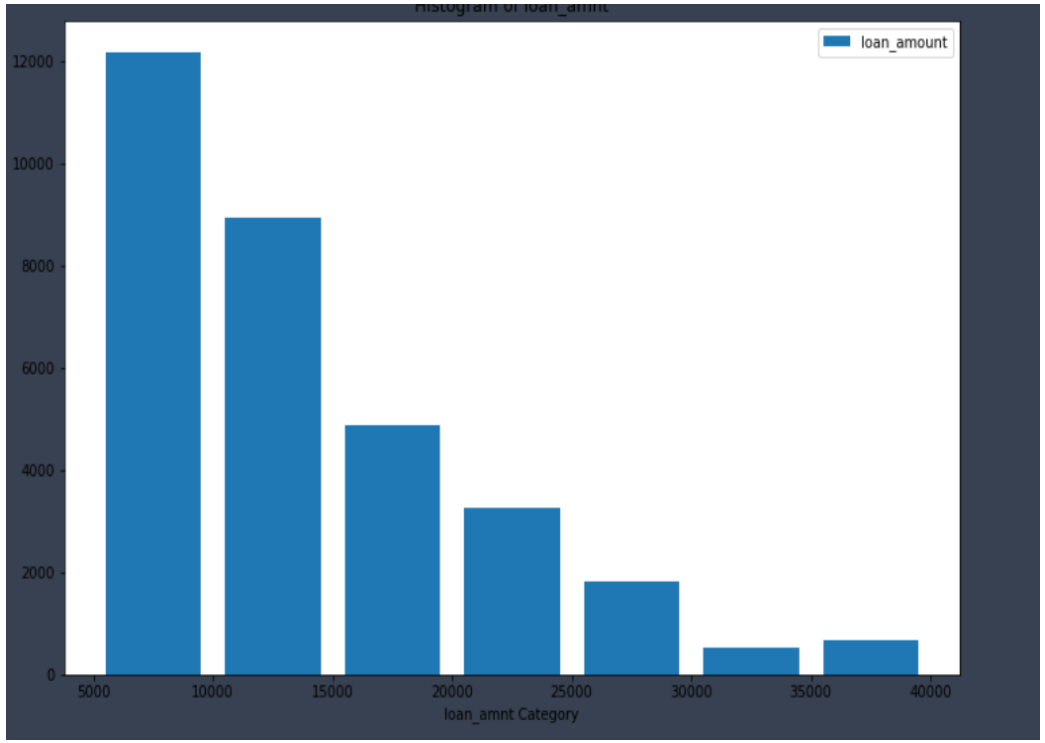
home_ownership: The home ownership status provided by the borrower during registration or obtained from the credit report. Our values are: RENT, OWN, MORTGAGE, OTHER

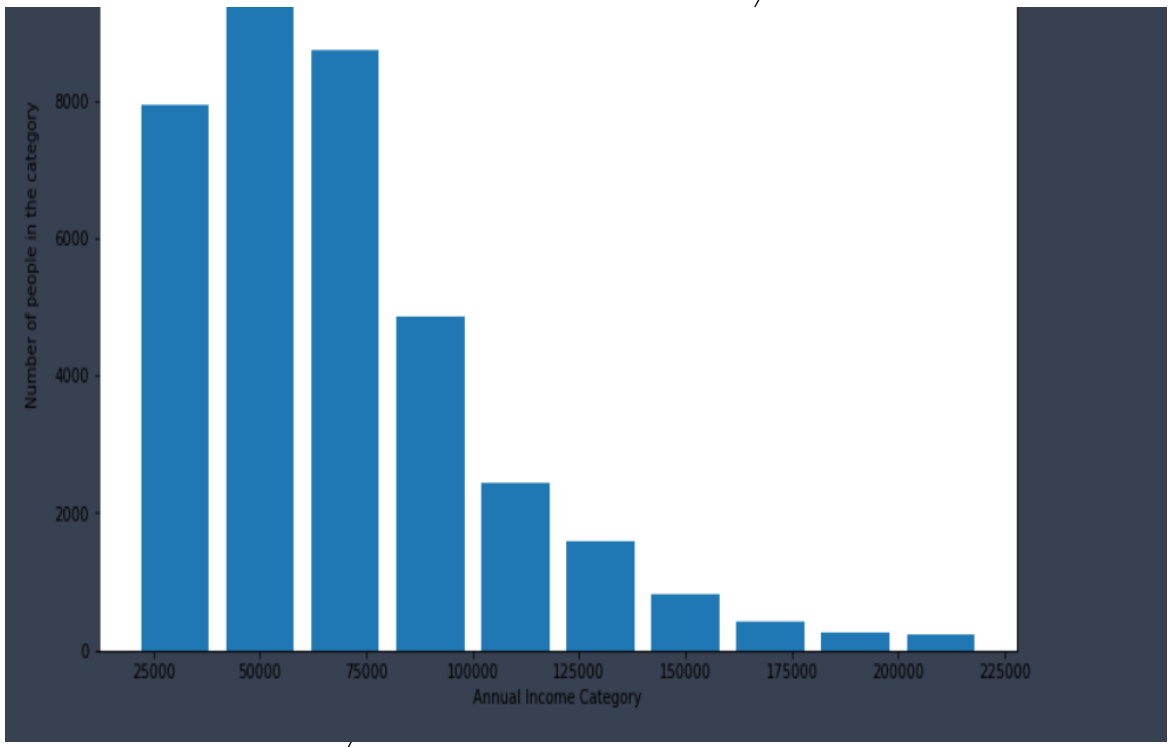
annual_inc: The self-reported annual income provided by the borrower during registration

Purpose: A category provided by the borrower for the loan request

SKEWNESS OF THE DATA

Most of the data shows that the data is right skewed which has to brought to bell curve.





NULL VALUES : NONE IN THIS DATASET

NULL VALUES Do not add any value during the prediction and indeed may cause issues to predict something which is right.

Hence as data scientist we tend to find the null values and handle them.

This data set do not have any null values.

OUTLIERS IN THE DATA AND CLEANING THEM.

From the skewness and from the below graph we will remove the extreme cases which is rare that the Lenderclub will however encounter. Thus to make sure we are considering the majority normal population, we found the outliers in the data and removed them.

We used the 3S rule to do so. Which says that most of the customers live within a 3 quartile range. We considered the quartile range of 25% to 75% to remove the outliers

FEATURE HANDLING

1. Encode the categorical values and convert them to number

2. Normalize the dataset to have them in the same scaling to improve the prediction performance

MODEL BUILDING

1. We divide the dataset into test and train and divide them to features and targets
2. Our target variable will be the interest_rate here which will be predicted by the features like the term of the loan, amount of the loan, and so on....
3. Since this is a continuous value we will use a regression model.
4. Famous regression models like Linear regression, Lasso and Decesion trees are used.
5. Have performed with various models to find the best fit model

RESULTS AND OBSERVATIONS

	model	best_score	best_params
0	linear_regression	0.897537	{'normalize': True}
1	lasso	-0.000395	{'alpha': 1, 'selection': 'random'}
2	decision_tree	0.851199	{'criterion': 'friedman_mse', 'splitter': 'ran...

The models were trained with different models and hypertuning pararmeters. Above are the results of the same

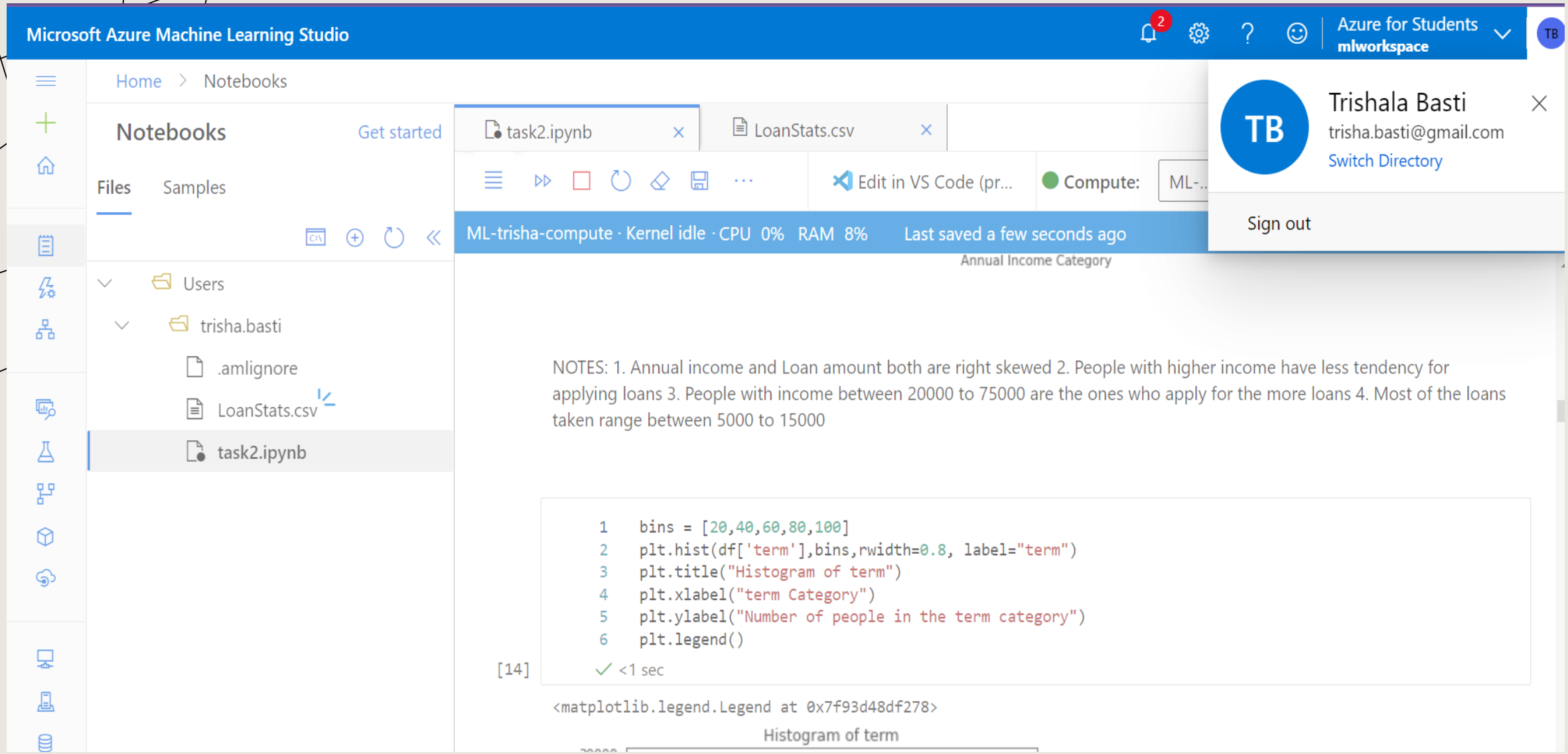
FINAL NOTE

1. The best score of the model is 89% which means the data is able to predict the interest rate on the test data 89% of the times.
2. The Model has the rmean square error reported as 0.89 which means it is good score again for the model.
3. The 0.89 of the RMSE also says the model is near to the overfitting but since it is not 1, it is not overfitting. Quite a good model for the dataset but more data was required to check this.
4. The RMSE score is not near to 0, hence not at all underfitting which is the best thing for a model.
5. The Linear regression is the best prediction model for this kind of data as the scores of the other model like lasso and decision trees is not upto this model

AZURE DEPLOYMENT

The model is not deployed in the azure cloud for future performance and MLOPS purpose

DEPLOYED IN AZURE



THANK YOU

TRISHALA BASTI

DATA SCIENTIS

Trishala.basti@student.htw-berlin.de