Loan Defaulter Prediction

Group 18 (Runtime Terror): Dishant Vakte, Jeffi Edelbert, Rakshit Sinha, Yatin Koul, Zhanyi Zhu, Zheng Cen

Introduction

- Defaulters could potentially cost banks a lot of revenue.
- Banks need a concrete way to judge the credibility of its future customers before issuing a credit card or a loan.
- Predicting if a customer will default or not, can be done based on various socio -economic factors.
- Determining these factors will help the bank forecast and filter out defaulters.

About The Dataset

- Direct marketing campaigns of a Portuguese banking institution; based on phone calls;
- 45211 instances and 17 attributes, total of 768,587 data points;
- Attributes like default, marital, job, education, housing loan, etc
- Highly imbalanced and requires resampling

Data Processing

Creating dummy variables for multiple categorical columns

Changing default variable to numerical column with binary values

Categorizing age column into bins

Detecting and removing outliers

Data Analysis

01

Socio-economic factors

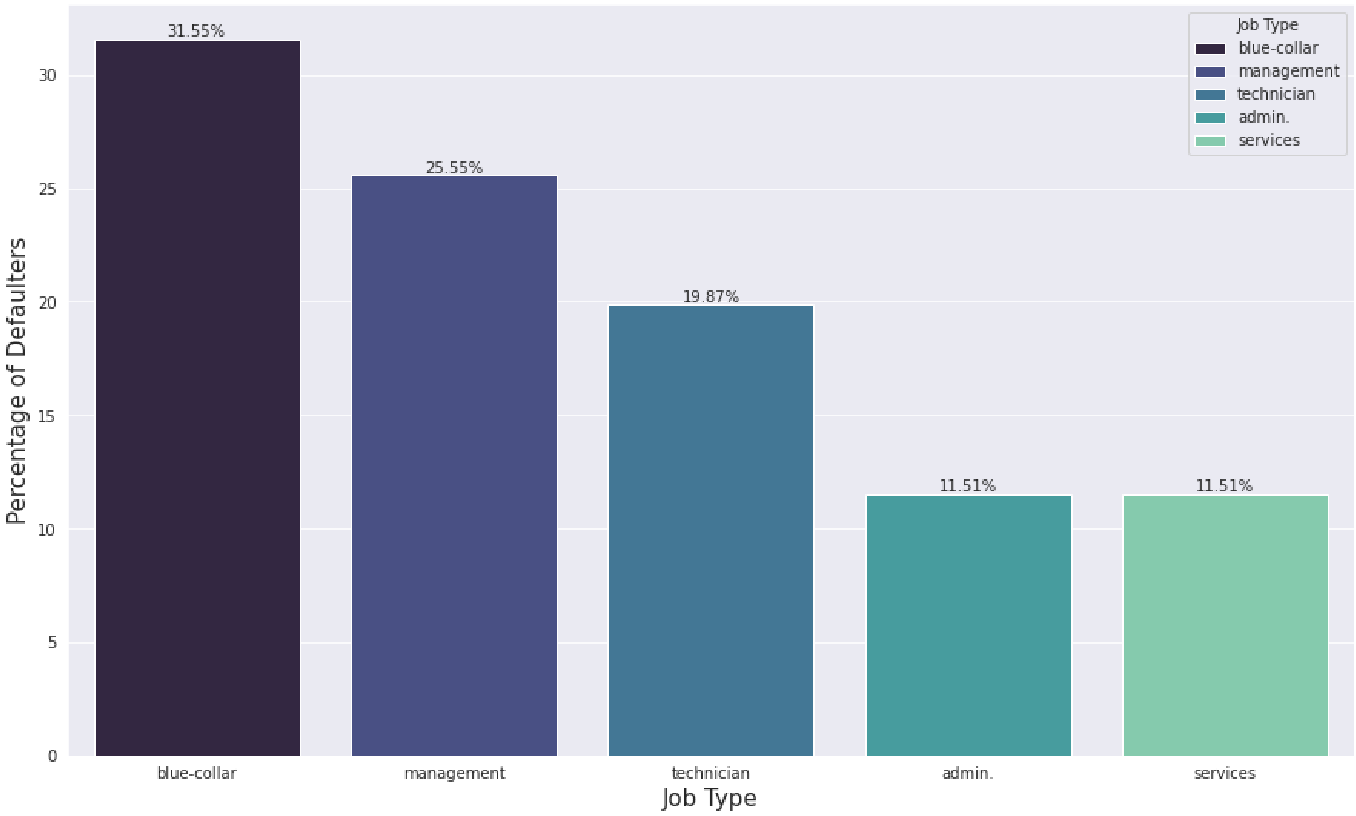
02

Correlation between variables

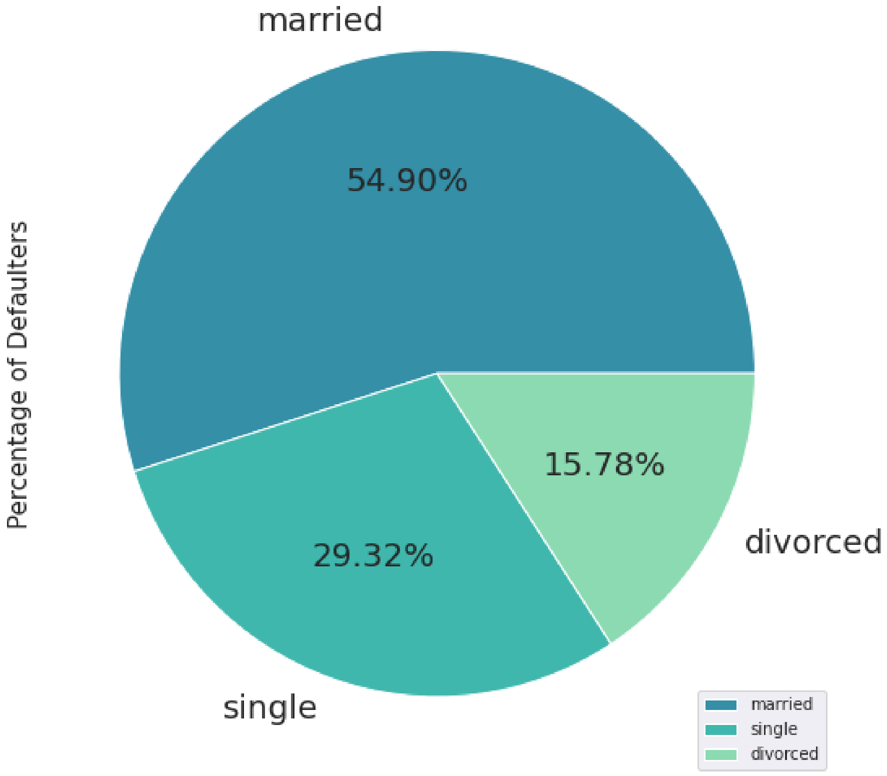
03

Machine Learning Models

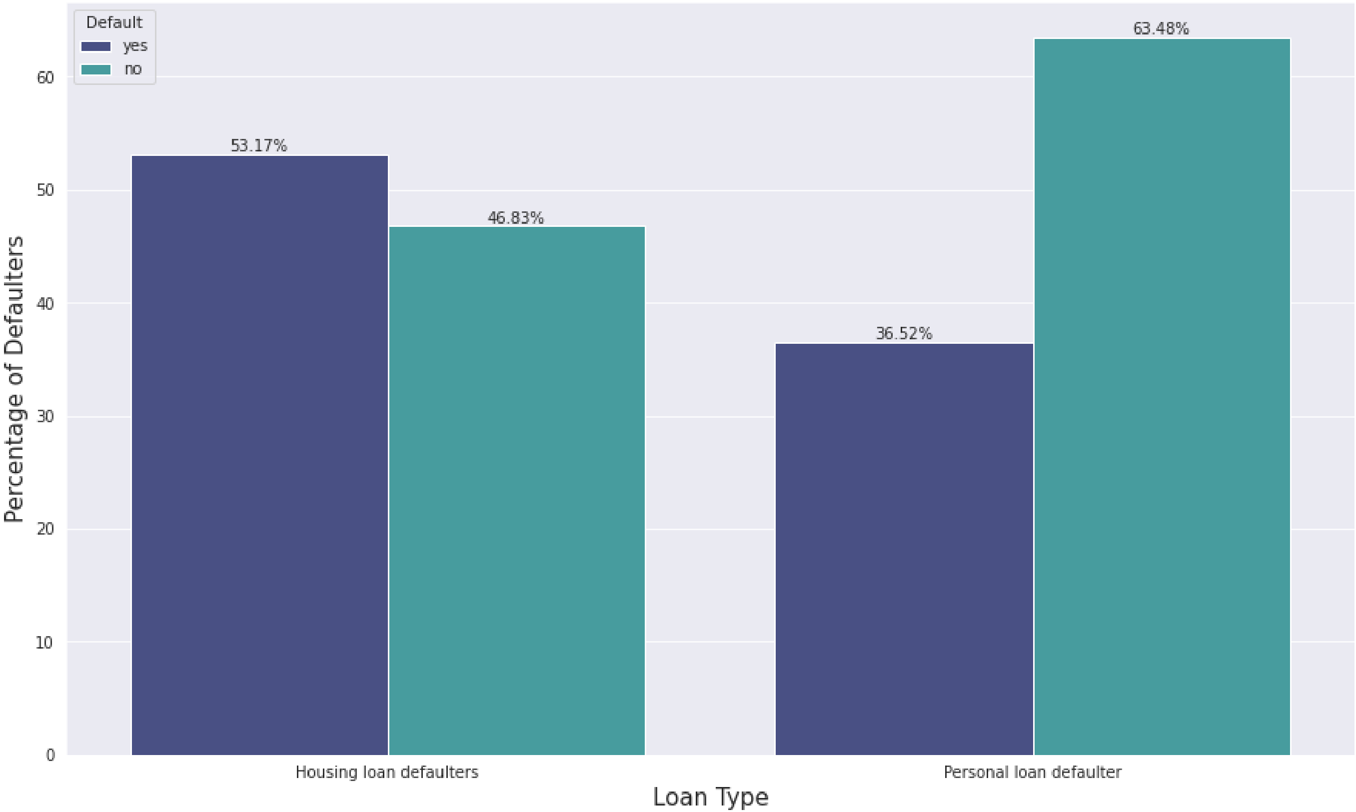
Socio-economic factors causing default - Job Type



Socio-economic factors causing default - Marital Status



Socio-economic factors causing default - Loan Type



Socio-economic factors causing customers to default

Marital Status

54.97% of defaulted customers are **married**.

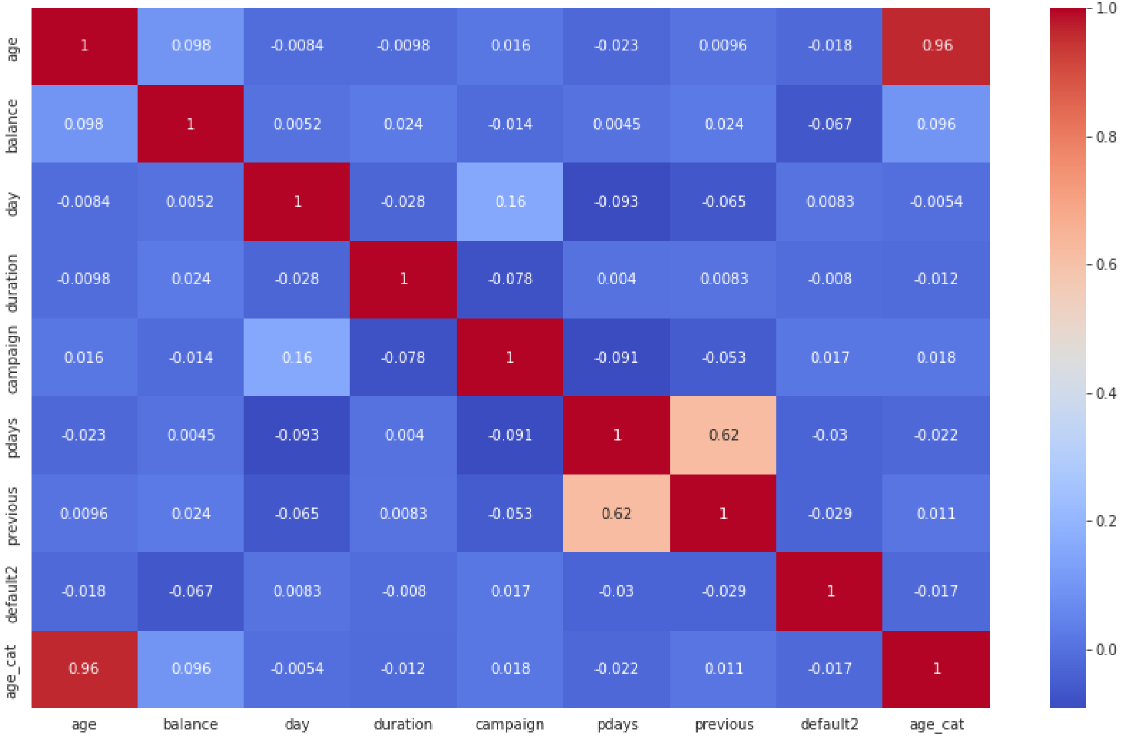
Job Type

31.21% of defaulted customers work a **blue -collar job**

Loan Type

53.37% of defaulters have a **housing loan**

Heatmap to analyze correlation between various parameters



Implementing Machine Learning Models

- → Machine Learning Algorithm: Random Forest
- → Re-sampling methods:
  - Undersampling
  - Oversampling
  - SMOTE

SMOTE

(Synthetic Minority Oversampling Technique)

- A hybrid model of undersampling as well as oversampling.
- Using data augmentation techniques, randomly generate new minority class data points.
- By this, we achieve high number of “unique” minority class samples.
- ML model will have a high number of samples to be trained on.

Random Forest

- An ensemble learning algorithm using a number of “Decision Trees”, which works on creating a split based on various parametric values.
- Random Forest is a good choice for highly imbalanced dataset
- Ensemble learning allocates “weights” or “importance” to the target class - high weight for minority class and less weight for majority class.
- Helps to build an efficient classification model.

Results

→ **Undersampling**
F-1 Score of **17%** on the original test dataset (scaled). Underfitting

→ **Oversampling**
F-1 Score of **0%** on the original test data (scaled). Severely overfitting

→ **SMOTE**
F-1 Score of **70%** on the original test dataset (scaled). Good performance.

Conclusion

- ➔ Customers falling under categories with high likelihood of defaulting such as **married**, **working blue-collar jobs** or **having a housing loan** could be offered higher interest rates.

- ➔ This would ensure reduced risks for the firm and greater caution towards these customers.

- ➔ **Recommendation:**
  This model can be built into an interface which allows executives to input information about a new customer and predict whether the customer is likely to default.

Thank You!