**Machine learning in the FinTech industry**

**Problem statement**: The company is facing challenges in the loan procedure, specifically the issue of defaulters. The company wants to transform its loan procedure using Machine Learning model to improve decision-making and prevent defaulters.

**Solution**: The company can use AI and automation to analyze customer data in real-time and predict the likelihood of a customer defaulting on their loan. Machine learning models such as logistic regression, decision trees, random forests, and neural networks can be used to make better decisions regarding loan approval. The data analytics team should consider relevant parameters such as having enough data to train the model accurately, choosing the right features to predict the likelihood of default, and considering the performance metrics to evaluate the model. The future of AI and automation in the loan procedure looks promising as it can improve decision-making, reduce the risk of defaulters, and provide better customer service. This can have a positive impact on the economy by improving the financial health of individuals and businesses, leading to increased investment and economic growth.

****

**how can our company transform with the use of AI and automation to solve the loan procedure and prevent defaulters?**

Overview of how we could implement this:

With the help of AI and automation, we can analyze customer data in real-time and predict the likelihood of a customer defaulting on their loan. We can use machine learning models such as logistic regression, decision trees, random forests, and neural networks to make better decisions regarding loan approval.

**What parameters should we consider when building these machine learning models?**

Data analytics team should consider several parameters when building these models. Firstly, we need to ensure that we have enough relevant data to train the model accurately. Secondly, we need to choose the right features to predict the likelihood of default. And finally, we need to consider the performance metrics such as precision, recall, and accuracy that we're using to evaluate the model.

**What do you think the future of AI and automation in the loan procedure looks like?**

I believe the future looks very promising for AI and automation in the loan procedure. As we continue to collect more data and improve our machine learning models, we'll be able to make better decisions regarding loan approval and reduce the risk of defaulters. Additionally, we can provide better customer service through automation and personalized loan servicing, which could lead to higher customer satisfaction and loyalty.

**How do you think this will affect the recession in the market?**

I think this will have a positive impact on the economy. By reducing the risk of defaulters, the financial health of individuals and businesses will improve, which could lead to increased investment and economic growth. It's a win-win situation for both our company and the economy.

**MLOps**

The MLOps pipeline can involve data collection and preprocessing, feature engineering, model training, model deployment, and monitoring. The pipeline can be automated using tools such as docker, Azure ML and Kubernetes. The pipeline can be deployed on cloud infrastructure such as AWS, GCP, or Azure.

****

**About Dataset**

A simulated financial dataset has been generated using genuine information from a financial organization. The dataset has been altered to eliminate any identifying characteristics and the figures have been altered to prevent any linkage to the original source (the financial institution). The purpose of using this dataset is to give trainee a simple financial dataset to use when practicing financial analytics for POC.

****

**Highlights of the Loan Default Classification:**

-Classification, Imbalanced Data, and PR Curve

**Contents**

1. EDA

* Check NaN values
* Data overview
* Feature engineering
* Data distribution

1. Modeling

* Train test split
* Standardization
* Upsampling by SMOTE
* Logistic regression
* Support vector machine
* Random forest
* LightGBM
* XGBoost
* Model assessment
* ROC curve
* PR curve

1. Conclusion

**Data overview**

Table

Description automatically generated

The column labeled "Employed" is of categorical type, while the "Bank Balance" and "Annual Salary" columns are numerical. Our objective is to perform a binary classification task based on the target column "Defaulted."

**Feature engineering**

Text

Description automatically generated

We generate a new feature named "Saving Rate" based on the "Bank Balance" and "Annual Salary" data. The Saving Rate feature provides insight into the spending habits of each user. Generally, a user with a higher Saving Rate is considered less likely to default. We will investigate the relationship between these variables in greater detail later on.

**Data distribution**

**Default distribution:**

Loan defaults would only impact 3% of customers, creating in an imbalanced classification.

Chart, pie chart

Description automatically generated

**Employed distribution:**

As their p-value is between 0.0005 and 0.05, we draw the conclusion that they are not independent. Employed status can therefore be used to predict default.

Chart, pie chart

Description automatically generated

Chart, sunburst chart

Description automatically generated

**Bank Balance distribution:**

Approximately 500 individuals have hardly saved any money in their bank accounts, which could pose a risk for loan defaults. Surprisingly, those who have defaulted on their loans tend to have a higher balance in their bank accounts. This observation may seem counterintuitive and suggests the presence of confounding factors. It is possible that individuals with a higher bank balance may have easier access to loans, leading to a higher number of defaults.

Chart

Description automatically generated

**Annual Salary distribution:**

In comparison to bank balance, there are fewer outliers when it comes to annual salary. Default cases appear to be distributed across all annual salary ranges, suggesting that annual salary may not be a reliable predictor of loan defaults.

Chart, histogram

Description automatically generated

**Saving Rate distribution**:

The distribution of saving rate is similar to that of bank balance, but with a few extreme outliers. This suggests that people's saving habits can vary significantly. Some individuals may earn a high income but spend more than they save, while others with relatively low salaries may have a significant amount of savings.

Graphical user interface

Description automatically generated with low confidence