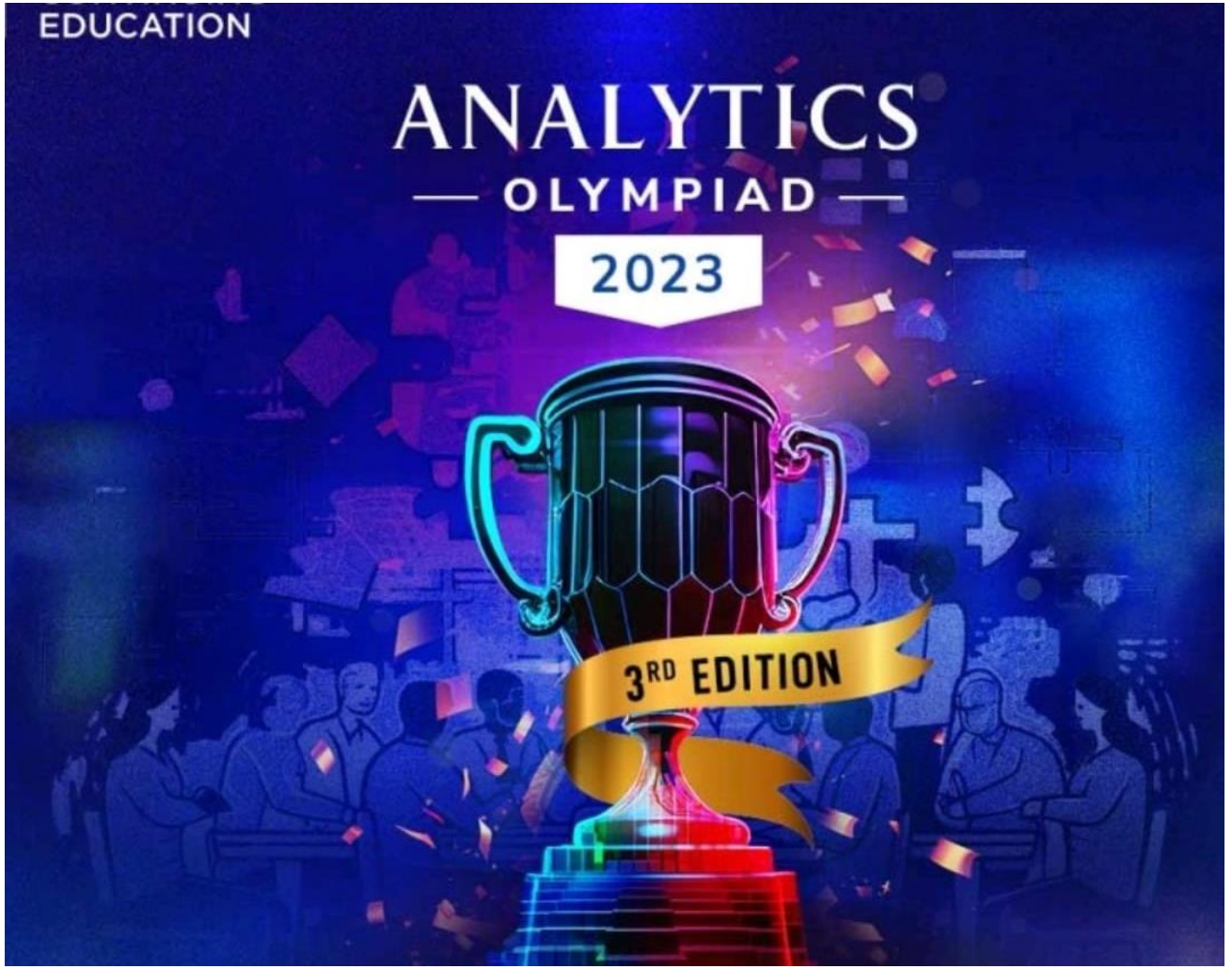
SHIV NADAR
INSTITUTION OF EMINENCE DEEMED TO BE UNIVERSITY
DELHI NCR

ACADEMY OF CONTINUING EDUCATION

MACHINE HACK

Predicting Customer Loan Default

Turning Data into Insights: Unveiling the Power of Ensemble Machine Learning



Presented by:- Aniket Bhausaheb Barphe

Manager-Data Science, India First Life Insurance Company

Analytics Olympiad 2023- MachineHack

https://www.linkedin.com/in/aniiketbarphe/

SHIV NADAR
INSTITUTION OF EMINENCE DEEMED TO BE UNIVERSITY
DELHI NCR

ACADEMY OF CONTINUING EDUCATION

MACHINE HACK

Outline

- Problem Statement
- Data Understanding
- Data Preparation
- Model Building & Evaluation
- Recommendations

SHIV NADAR
INSTITUTION OF EMINENCE DEEMED TO BE UNIVERSITY
DELHI NCR

ACADEMY OF CONTINUING EDUCATION

MACHINE HACK

Problem Statement

- Analytics Olympiad 3.0, organized by the Academy of Continuing Education at Shiv Nadar Institution of Eminence in collaboration with MachineHack, centers around the theme of "Predicting Customer Loan Default." This competition is designed to offer valuable insights to the banking, finance, insurance (BFSI), and fintech industries. In today's dynamic financial landscape, the ability to accurately predict customer loan default is crucial for mitigating risks and maintaining a robust lending portfolio.

- Participants in this challenge are tasked with leveraging machine learning techniques to develop predictive models. These models aim to assess the likelihood of a customer defaulting on a loan. To make these predictions, participants will utilize data related to the customer's credit history, payment behavior, and account details. By doing so, they contribute to the broader goal of enhancing risk assessment and ensuring the financial well-being of lending institutions.

Data Understanding

In-Depth Data Overview:-

1. **Train Data Dimensions:** The training dataset consists of 1,106,674 rows and 62 columns, showcasing the breadth of data available for analysis.
2. **Test Data Dimensions:** Our test dataset comprises 474,289 rows and 60 columns, ready for evaluation and prediction.
3. **Categorical Features Transformation:** Categorical features have been effectively transformed into numerical values, addressing missing values in both the training and test datasets.
4. **Multi-Label Classification Problem:** With two target variables in focus, our task is categorized as a Multi-Label Classification problem. This necessitates specialized techniques to effectively predict and interpret multiple target outcomes.

Data Preparation

Detailing Data Cleaning and Preprocessing for Model Building:-

1. **Handling Missing Values:** Addressing missing values in encoded features of categorical variables, we replaced them with the mode (most frequent value) for both the training and test datasets.
2. **Outlier Analysis:** We experimented with the 'Z score' method to identify and potentially remove outliers. However, it did not yield a significant improvement in model performance, leading us to exclude 'Outlier removal' from our final analysis
3. **Feature Engineering:** To enhance model efficiency and mitigate the effects of correlation among existing features, we created 13 new features. The details and definitions of these new features are available in the shared Jupyter notebook.
4. **Feature Selection Process:**

- **Initial Model Testing:** We began by testing models using all original features (excluding derived features, cust_id, first name, and last name). The accuracy score ranged between 0.50 to 0.60.
- **Inclusion of Derived Features:** In the second phase, we incorporated all features (both original and derived, excluding cust_id, first name, and last name). The accuracy score improved to a range of 0.55 to 0.68.
- **Correlation Analysis:** After analyzing the correlation matrix, we finalized a set of 30 features for building the model, which included a combination of original and derived features. The list of these final features can be found in the shared Jupyter notebook.

5. **Addressing Class Imbalance:** Notably, due to a class imbalance issue, we chose not to split the training dataset for validation purposes. Instead, we utilized the entire training dataset to ensure the model was trained on a maximum number of records

SHIV NADAR
INSTITUTION OF EMINENCE DEEMED TO BE UNIVERSITY
DELHI NCR

ACADEMY OF CONTINUING EDUCATION

MACHINE HACK

Model Building & Evaluation

Detailed Overview of Models Used and Observed Results:-

1. **Differential Modeling Approach:** One notable aspect of our modeling strategy is the separate prediction of each target variable, utilizing distinct model configurations. This approach allowed us to identify whether hyperparameter tuning was necessary for each model, excluding the base model, which remained consistent for both target variables.
2. **Diverse Model Set:** We employed a total of 25 models, encompassing both base models and versions with hyperparameter tuning.
3. **Performance Highlights:** Among these 25 models, select models, such as GradientBoost, XGBoost (XGB), and CatBoost, demonstrated superior predictive capabilities for both target variables.
4. **Detailed Model Performance:** The performance metrics of the remaining models, especially on the test dataset, are available for reference in our Jupyter notebook.
5. **Feature Consistency:** It is worth noting that we did not observe significant variations in feature importance across different models. Despite experimenting with various feature combinations for each model, the utilization of the previously selected 30 features, as outlined in the data preparation for model steps, consistently delivered strong results.
6. **Recommended Deployment Model:** For deployment purposes, we recommend 'Model 25,' which is an ensemble version combining the strengths of the Gradient Boosting Classifier with optimized hyperparameters and the base version of the XGBoost (XGB) Classifier. This ensemble model reduces deviation in individual models, resulting in robust predictions.

Recommendations

Summary of Innovative Ideas:-

1. **Multi-Label Classification Approach:** We addressed the challenge as a Multi-Label Classification problem, simultaneously predicting two target variables. This innovative approach captures intricate relationships between variables, enhancing predictive capabilities.
2. **Feature Engineering for Efficiency:** The creation of 13 new features was a strategic move to enhance model efficiency and reduce feature correlation's impact on predictions.
3. **Iterative Feature Selection:** Our iterative feature selection process, based on correlation analysis, resulted in a set of 30 key features that significantly contributed to model performance.
4. **Class Imbalance Handling:** We adopted a creative approach to handle class imbalance by using the entire training dataset for model training, ensuring maximum exposure to different records.

These innovative strategies and deployment recommendations demonstrate our commitment to delivering reliable, data-driven solutions for predictive modeling in the context of loan default prediction. Our methods offer valuable insights for deployment in the BFSI and fintech sectors.

Recommendations(Cont..)

Business Outcomes and Real-Life Impact:-

1. **Enhanced Risk Assessment:** The predictive models developed in this analysis offer valuable tools for financial institutions, particularly in the banking, finance, insurance (BFSI), and fintech sectors. By accurately assessing the likelihood of customer loan default, these institutions can make more informed lending decisions.
2. **Mitigated Risks:** Accurate predictions of customer loan default enable lending institutions to mitigate risks effectively. This means fewer non-performing loans, reduced financial losses, and a healthier lending portfolio.
3. **Financial Well-being:** The analysis directly contributes to the financial well-being of lending institutions by improving their ability to manage loan defaults. This, in turn, safeguards their sustainability and profitability in the dynamic financial landscape.

In summary, this analysis not only offers practical solutions for predicting loan defaults but also showcases innovative techniques in data preprocessing, feature engineering, and model selection. These strategies have the potential to significantly impact the financial industry by improving risk assessment and lending portfolio management.

SHIV NADAR
INSTITUTION OF EMINENCE DEEMED TO BE UNIVERSITY
DELHI NCR

ACADEMY OF CONTINUING EDUCATION

MACHINE HACK

THANK YOU FOR YOUR ATTENTION