**Problem Statement: Predicting Credit Risk for Loan Applicants**

**Background:**

Financial institutions face significant challenges in assessing the creditworthiness of loan applicants. Accurate credit risk prediction is crucial for minimizing defaults and ensuring the stability of the lending system. The German Credit dataset provides a comprehensive set of features related to applicants' financial history, personal information, and loan details, making it an ideal resource for developing predictive models.

**Objective:**

The objective of this project was to develop a machine learning model to predict the credit risk of loan applicants using the German Credit dataset. The model aims to classify applicants into two categories: good credit risk and bad credit risk. Additionally, insights into the key factors influencing credit risk were provided, and strategies for improving the credit evaluation process were suggested.

**Requirements:**

**1. Data Exploration and Preprocessing:**

* **Analyze the dataset**: The dataset consists of various features like applicant's credit history, job type, age, and more, which were explored for insights.
* **Handling missing values**: Missing values were handled appropriately to avoid inaccuracies in the model.
* **Feature engineering**: Several features were created or modified to improve model performance.

**2. Model Development:**

* **Machine learning algorithms**: Logistic Regression, Decision Tree, Random Forest, XGBoost, LightGBM, and SVM were tested for classification.
* **Model training and evaluation**: Each model was trained and evaluated using metrics such as accuracy, precision, recall, F1-score.
* **Hyperparameter tuning**: Techniques like grid search and cross-validation were applied to improve model performance.

**3. Model Interpretation and Insights:**

* **Model performance**: Various models were compared based on their performance in predicting credit risk.
* **Feature importance**: Key features influencing the predictions were identified using the trained models.
* **Visualizations**: Visual aids were used to represent the results and provide better insights into model performance and data characteristics.

**Methodology:**

**Data Exploration and Preprocessing:**

* **Initial Data Check**: The German Credit dataset was analyzed to understand the distribution of features and target variables.
* **Missing Data Handling**: Missing values were handled by either dropping rows with missing data or using imputation techniques.
* **Feature Scaling**: Features were standardized to ensure they contribute equally to model performance.

**Model Development:**

* **Logistic Regression**: A baseline model was built using logistic regression to predict the credit risk.
* **Decision Tree**: A decision tree was trained to identify decision boundaries and interpret decision paths.
* **Random Forest**: The random forest model, an ensemble technique, was used for better generalization by combining multiple decision trees.
* **XGBoost**: XGBoost, known for its performance in structured data, was used and tuned for optimal performance.
* **LightGBM**: Another gradient boosting model, LightGBM, was tested due to its efficiency in large datasets.
* **SVM**: Support Vector Machine was tested for classification due to its robustness in high-dimensional spaces.

**Model Tuning:**

* **Hyperparameter Optimization**: Various hyperparameters like learning rate, number of estimators, and tree depth were tuned using grid search.
* **Cross-validation**: Cross-validation was used to ensure the robustness and generalizability of the models.

**Results:**

**Evaluation Metrics:**

Below are the evaluation results for various models on the dataset:

1. **Logistic Regression**:
   * Accuracy: 78%
   * Precision: 93% (class 1), 44% (class 0)
   * Recall: 78% (class 1), 74% (class 0)
   * F1-Score: 85% (class 1), 55% (class 0)
2. **Decision Tree**:
   * Accuracy: 77%
   * Precision: 91% (class 1), 43% (class 0)
   * Recall: 79% (class 1), 68% (class 0)
   * F1-Score: 85% (class 1), 53% (class 0)
3. **Random Forest**:
   * Accuracy: 82%
   * Precision: 89% (class 1), 56% (class 0)
   * Recall: 90% (class 1), 53% (class 0)
   * F1-Score: 90% (class 1), 54% (class 0)
4. **XGBoost**:
   * Accuracy: 82%
   * Precision: 88% (class 1), 55% (class 0)
   * Recall: 91% (class 1), 45% (class 0)
   * F1-Score: 89% (class 1), 49% (class 0)
5. **LightGBM**:
   * Accuracy: 84%
   * Precision: 87% (class 1), 62% (class 0)
   * Recall: 94% (class 1), 42% (class 0)
   * F1-Score: 90% (class 1), 50% (class 0)
6. **SVM**:
   * Accuracy: 81%
   * Precision: 87% (class 1), 60% (class 0)
   * Recall: 94% (class 1), 39% (class 0)
   * F1-Score: 90% (class 1), 48% (class 0)

**Conclusion:**

Based on the evaluation metrics, **LightGBM** performed the best with an accuracy of **84%** and a high recall rate for class 1 (good credit risk). This model balances the need for high recall (minimizing false negatives) and high precision (minimizing false positives), making it the most suitable for the credit risk prediction task.

**Recommendations**:

* **Feature Engineering**: Additional domain-specific features could be added to further improve model performance.
* **Data Collection**: Collecting more data for applicants with bad credit risk could help improve the model's precision for this class.
* **Model Monitoring**: Regular monitoring of the model in a real-world application should be conducted to ensure its continued relevance and accuracy.

**GitHub Link:**

You can view the complete code and results on GitHub using the following link: [TCS Hackathon GitHub Repository](https://github.com/Borazonic/TCS-Hackathon/blob/main/TCS%20Hackathon.ipynb)