Data Intensive Computing

Project Phase 2

Name	UBIT Name	Person Number	Contribution (in %)
Chirayu Sanghvi	chirayus	50545042	33.33 %
Jayant Sohane	jsohane	50533812	33.33 %
Naveen Veeravalli	nveerava	50496337	33.33 %

Enhancing Credit Default Prediction for Improved Lending Practices

1. Problem Statement

The challenge of accurately predicting credit card defaults is a fundamental concern in consumer lending. The project aims to address this issue, by developing machine learning models, capable of outperforming existing methods, thereby enhancing risk assessment practices and ultimately improving the learning experience for both financial institutions and customers.

Background:

Credit card default prediction is pivotal to the sound operation of consumer lending businesses. It determines whether the borrowers will repay their credit card balances promptly which, in turn, influences lending decisions, customer experiences and financial stability. Accurate prediction, mitigate risk, minimize loss, and foster more efficient lending processes. The significance of this problem lies in its direct impact on the profitability, sustainability, and customer satisfaction of lending institutions. Defaults can result in substantial financial losses and, in some cases, even threaten the viability of these institutions.

Potential Contributions:

1. Optimize lending decisions: Enhanced predictive models enable lending institutions to make better informed decisions regarding credit approvals and lending limits. This contributes to improved risk management and reduced default rates.

2. Improved customer experience: More accurate predictions can lead to fairer credit decisions, resulting in increased approval rates for credit card applications. This in turn provides customers with greater access to financial services and a smoother application process.

3. Enhanced Financial Stability: Lower default rates and better risk management can have a direct positive impact on the financial stability of lending institutions. Reduced defaults translate into reduced losses and improved profitability.

4. Foster Innovation: The development of advanced machine learning model encourages innovation within the lending industry. This project's findings can inspire further research and technological advancements in credit risk assessment.

5. Strengthen Industry competitiveness: Lending institutions that adopt more effective credit default prediction models gain a competitive edge in the market. This can lead to improved market share, and sustained growth.

Results and Analysis

- Performance Metrics for all model we applied on our model:

Parameter	Accuracy	Precision	Recall	f1 score
Logistic Regression	0.97914	0.94581	0.78796	0.85970
Support Vector Machine	0.98136	1.00000	0.77018	0.87017
XGBoost	0.98136	1.00000	0.77018	0.87017
Random Forest Classifier	0.97992	0.94354	0.80027	0.86602
Decision Tree Classifier	0.97437	0.83512	0.85227	0.84360
KNN Classifier	0.98141	0.97859	0.96701	0.85233

- Discuss the effectiveness of the algorithm when applied to your data to answer questions related to your problem statement.

Model	Pros	Cons
Logistic Regression	1) High Overall Accuracy	1) Relatively lower recall compared to precision, indicating a higher false negative rate. May miss some instances of credit card defaults.
Support Vector Machine	1) Perfect Precision, indication no false positive. 2) High Overall Accuracy	1) Lower Recall compared to precision, suggesting a potential for false negatives. 2) It may not capture all instances for credit card defaults.
XGBoost	1) Perfect Precision, high overall accuracy.	1) Similar to SVM, lower recall compared to precision. 2) It may not capture all the instances for credit card defaults.
Random Forest Classifier	1) High accuracy, good precision, and recall.	1) Slightly lower precision compared to SVM and XGBoost. 2) It may have a slightly higher false positive rate.
Decision Tree Classifier	1) Balanced precision and recall	1) Lower precision compared to SVM and XGBoost. May have a higher false positive rate.
KNN Classifier	1) High accuracy, precision, and recall.	1) KNN models can be computationally expensive, especially with large

		datasets. It might not be as interpretable as some other models.

Additional Observations:

- Computational Complexity: SVM and XGBoost are generally computationally expensive compared to simpler models like Logistic Regression or Decision Trees. Consider the computational resources available.

- Interpretability: Decision Trees are inherently more interpretable than ensemble models like Random Forest or complex models like SVM and XGBoost. If interpretability is crucial, this should be a factor in your decision.

- Scalability: Random Forest, XGBoost and SVM may perform well on large datasets, but their scalability might be a concern. Logistic Regression and Decision Trees are typically more scalable.

In summary, the choice of the best model depends on our specific needs and constraints. For interpretability, logistic regression and decision tree might be a good choice. On another note, for prioritizing precision and if we can handle computational complexity, SVM, XGBoost, or KNN could be suitable.

- Work we had to do to tune/ train the model:

Model	Work we did to tune this model
Logistic Regression	1) Using L1 Regularization (‘penalty=’l1’) for feature selection. 2) Choosing the ‘liblinear’ solver, suitable for our dataset. 3) Setting the regularization strength parameter (‘C’) to 10. 4) Properly splitting data, training the model, making predictions, and evaluating with metrics. 5) Visualizing metrics

Support Vector Machine	1) We have created a linear SVM classifier, 'SVC (kernel='linear')' 2) After that we have trained the model using train data and made prediction on test data.
XGBoost	1) Number of Estimators('n_selections'): set to 15 for boosting rounds. 2) Random State('random_state'): Fixed at 42 for reproducibility. 3) Scale Positive Weight('scale_pos_weight'): Adjusted based on 'classRatio' for handling imbalanced classes.
Random Forest Classifier	1) Number of estimators ('n_estmators'): Set to 50 for the number of decision trees. 2) Cost-Complexity Pruning ('ccp_alpha'): Set to 0.00001 for pruning control to control tree complexity. 3) Class Weight('class_weight'): Specified based on 'class_weight' for handling imbalanced classes. 4) Random State('random_state'): Fixed at 42 for reproducibility.
Decision Tree Classifier	1) Cost-Complexity Pruning ('ccp_alpha'): Set to 0.00001 for pruning control to control tree complexity. 2) Class Weight('class_weight'): Specified based on 'class_weight' for handling imbalanced classes. 3) Random State('random_state'): Fixed at 42 for reproducibility.
KNN Classifier	1) Number of Neighbors ('n_neighbors=5')

- Models we applied to determine the performance of our model on unseen data and why did we choose this model:

1) Logistic Regression:

We have used logistic regression on our model because of the following reasons:

- Simplicity and Interpretability:
    - Logistic regression provides a transparent and interpretable model. This simplicity is valuable in the context of financial institutions and customers, where understanding the model's decisions is crucial for regulatory compliance and improving the learning experience.

- Linearity and Feature Importance:
    - Despite assuming a linear relationship, logistic regression can capture significant linear relationships between independent variables and credit card defaults. This aids in identifying critical factors contributing to defaults, enhancing risk assessment practices.

- Handling Class Imbalance:
    - Credit card default datasets often exhibit class imbalance, where defaults are a minority class. Logistic regression can handle this imbalance through class weighting or sampling techniques., ensuring a balanced approach to risk assessment.

- Probability Estimation:
    - It directly models the probability of credit card defaults occurring. This probability estimation can assist financial institutions in setting specific risk thresholds and making informed decisions.

- Regularization:
    - This model can be regularized using L1 or L2 penalties to prevent overfitting, maintaining a balance between model complexity and generalization, which is crucial in improving risk assessment practices.

- Low Computation Cost:
    - This is computationally efficient and scales with large datasets, making it practical for enhancing risk assessment practices in consumer lending.

- Proven Performance:
    - Logistic regression has a track record of success in various classification tasks, including credit card default prediction, making it a reliable choice for outperforming existing methods, and improving risk assessment in this context.

b) Effectiveness of our model:

2) Support Vector Machine:

We have used Support Vector Machine on our model because of the following reasons:

- Non-Linearity Handling:
    - SVMs can capture complex, non-linear relationships in credit card default data, enhancing predictive accuracy.

- High-Dimensional Data:
    - SVMs are effective in dealing with high-dimensional datasets common in credit card default prediction.

- Margin Maximization:
    - SVMs maximize the margin between classes, providing more robust and confident risk assessment.

- Outlier Robustness:
    - SVMs are less sensitive to outliers, ensuring stable predictions in real world.
    - scenarios.
- Class Imbalance Handling:
    - SVMS can manage class imbalance, a common challenge in credit card default datasets.

- Interpretability:
    - SVMs provide transparent, interpretable results, crucial for regulatory compliance and learning experience improvement.

- Effective Regularization:
    - SVMs use regularization parameters to prevent overfitting and ensure reliable risk assessment.

3) XGBoost:

We have used XGBoost on our model because of the following reasons:

- Exceptional Predictive Performance:
    - This model is renowned for its outstanding predictive accuracy, a critical factor in enhancing risk assessment practices for credit card defaults, benefiting both financial institutions and customers.

- Handling Non-Linearity:
    - Credit card default prediction often involves complex, non -linear relationships between predictors. XGBoost excels at capturing these non-linear patterns, contributing to the project's goal of improving prediction accuracy.

- Feature Importance:
    - XGBoost provides feature importance scores, facilitating the identification of key factors influencing credit card defaults, a crucial component of risk assessment practices.

- Regularization:
    - XGBoost supports L1 and L2 regularization, allowing control over model complexity and prevention of overfitting, thereby enhancing generalization.

- Handling missing values:
    - XGBoost's ability to handle missing values without requiring imputation enhances robustness in cases of incomplete data, contributing to the project's effectiveness.

- Class Imbalance Handling:

    - Techniques such as weighted loss functions and subsampling, aid in addressing the inherent class imbalance often present in credit card default datasets.

- Fast Training and Predictions:
    - The computational efficiency and optimized speed of XGBoost make it a practical choice,
    - aligning with the need for swift credit card default prediction.

- Cross-Validation and Hyperparameter Tuning:
    - XGBoost's built in support for cross-validation and automated hyperparameter tuning simplifies the process of optimizing the model configuration for the project's objectives.

4) Decision Tree Classifier:

Using Decision Trees for Enhanced Credit Card Default Prediction because of the following reason:

Accurate credit card default prediction is crucial for the financial industry. In this context, decision trees offer several key advantages:

- Transparency: Decision trees provide clear, understandable models, ensuring transparency and accountability, vital in financial applications

- Identifying key factors: Decision trees help pinpoint the most critical factors contributing to credit card defaults, aiding in assessment.

- Handling Complex Relationships: They can capture complex, non-linear relationships in the data, a common occurrence in default prediction.

- Handling Missing Data: Decision trees handle missing data without the need for extensive data filling.

- Scalability: They efficiently process large datasets, making them suitable for real-time default prediction.

- No Linearity Assumption: Decision trees don't assume linear relationships, making them flexible.

- Improved accuracy with Ensembling: Combining decision trees through ensembles can enhance predictive accuracy, mitigating overfitting.

- Mixed Data Types: They handle both categorical and numerical features without extra transformations.

- Class Imbalance Handling: Effective in dealing with the imbalance between default and non-default cases.

- Interactive Adaptation: Decision trees are user-friendly and can be fine-tuned as needed.

- Regulatory Compliance: Their transparency ensures adherence to regulatory requirements.

- Efficiency: Decision trees have low computational overhead, suitable for resource-constrained environments.

5) Random Forest Classifier:

Enhancing Credit Card Default prediction with Random Forest classifier for the following algorithm:

Accurate credit card default prediction in vital in consumer lending. For this purpose, Random Forest offers compelling advantages:

- High Predictive Accuracy: By combining multiple decision trees, it improves accuracy and reduces overfitting in predicting credit card defaults.

- Interpretability: Despite being an ensemble model, it provides feature importance, helping understand default risk drivers.

- Handling non-linearity: It captures complex, non-linear relationships common in default prediction.

- Handling missing values: It deals with missing data effectively, enhancing robustness.

- Robust to Outliers: Random Forest is less affected by outliers often present in default datasets.

- Scalability: It efficiently handles large datasets, suitable for real-time processing.

- Versatility: Suitable for various aspects of credit card risk assessment, supporting both classification and regression tasks.

- Class Imbalance Handling: It addresses class imbalance effectively, vital for predicting credit card defaults.

- Low Overfitting: Aggregating results from multiple trees mitigates overfitting, ensuring stable predictions.

- Cross-Validation and Hyperparameter Tuning: Supports optimizing model configuration for enhanced performances.

- Reduction in variance: Improves model robustness and stability.

- Active Development and Support: Maintained with a strong community, ensuring reliability.

Random aligns with the project's aim of enhancing risk assessment in credit card default prediction.

6) K-nearest Neighbors

Enhancing Credit Card Default Prediction with K-nearest Neighbors (KNN) because of following reasons:

Accurate credit card default prediction is crucial in consumer lending. KNN offers key advantages:

- Simplicity: Easy to understand and implement, prioritizing simplicity and interpretability.
- Non-parametric: Adapts to diverse data types, including non-linear relationships, without making distribution assumptions.
- Handling non-linearity: Effectively captures non-linear patterns without complex model specifications.
- No Training Phase: Adaptable to dynamic datasets where patterns change over time.
- Handling Outliers: Robustness to outliers with adjustable neighbor count (k).
- Handling Missing Values: Can impute missing data based on nearest neighbors, enhancing robustness.
- Class Imbalance Handling: Naturally addresses class imbalance through weighted neighbors.
- Local Patters: Suits scenarios with varying credit card default risk factors within dataset segments.

- Ensemble Learning: it can improve performance through ensemble frameworks.

- Adaptability: Adapts to changing data distributions, valuable for evolving default nature.

- Cross-Validation and Hyperparameter Tuning: Optimizable for the credit card default prediction task using these techniques.

While KNN is viable choice when simplicity and interpretability are prioritized over predictive performance.
