PROBLEM 1:

A car manufacturer known for making large automobiles is struggling with sales and has asked for your help in designing an energy efficient car. Using data gathered, determine which attributes may contribute to higher gas mileage so that they can design a more fuel-efficient automobile.

Part 1:

Use proper data cleansing techniques to ensure that you have the highest quality data to model this problem. Detail your process and discuss the decisions you made to clean the data.

Part 2:

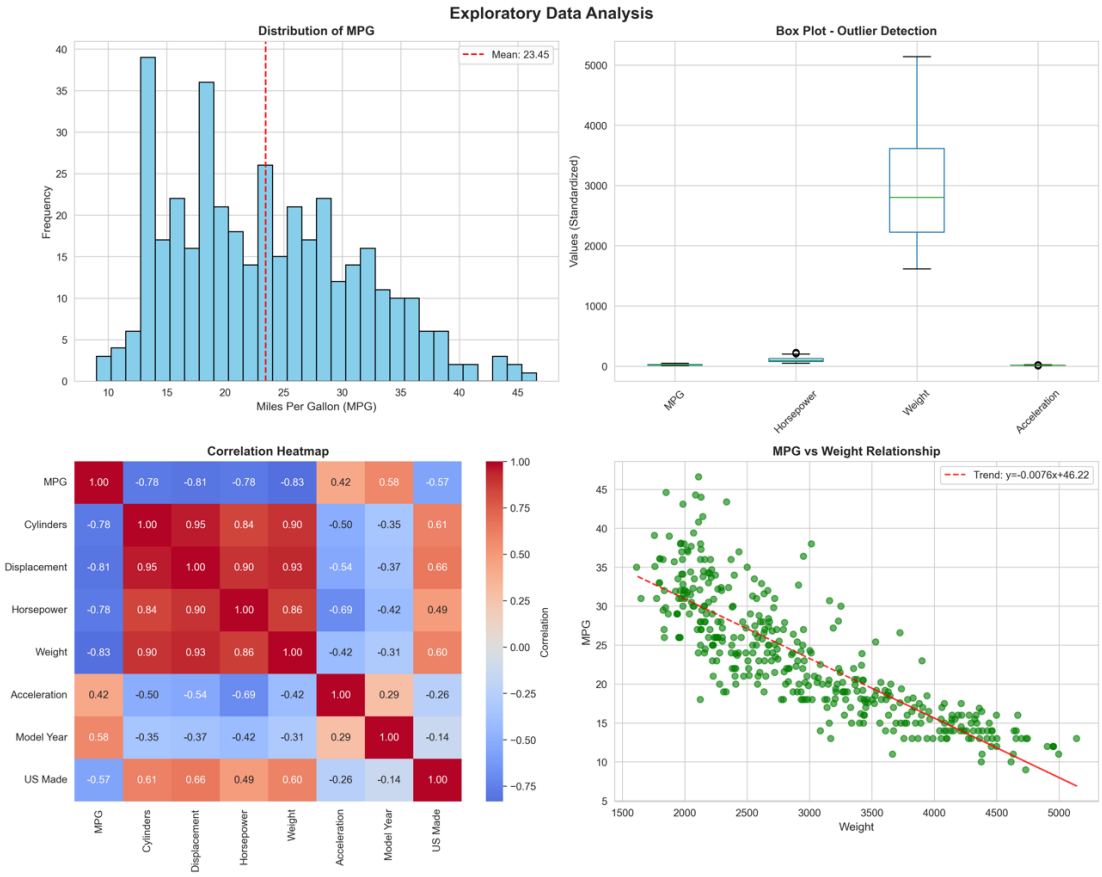
Build a linear regression model to accurately predict miles per gallon (MPG) based on the attributes of a vehicle. Discuss the significant attributes and how they can help you build the proper car.

Part 3:

Optimize the model using selection techniques, explain whether the model can achieve the specified goals, and describe which attributes contribute to higher MPG over others.

PROJECT 1: Building the car of the future: a data-driven approach to fuel efficiency

PROJECT TITLE: Predictive Model for Vehicle Fuel Efficiency



CHALLENGE

- Automobile manufacturer needed to identify which vehicle
- characteristics drive fuel efficiency to inform redesign strategy.

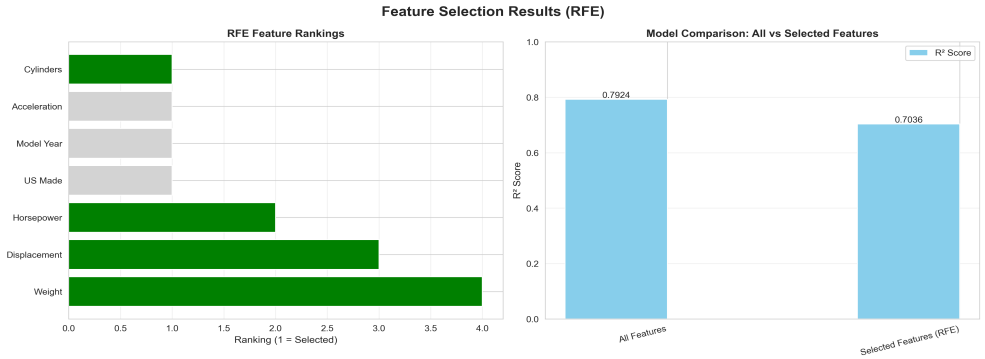
MY APPROACH

- Cleaned 398-record dataset, handling missing values
- Applied multiple linear regression using Python
- Optimized model using Recursive Feature Elimination
- Translated findings into business recommendations

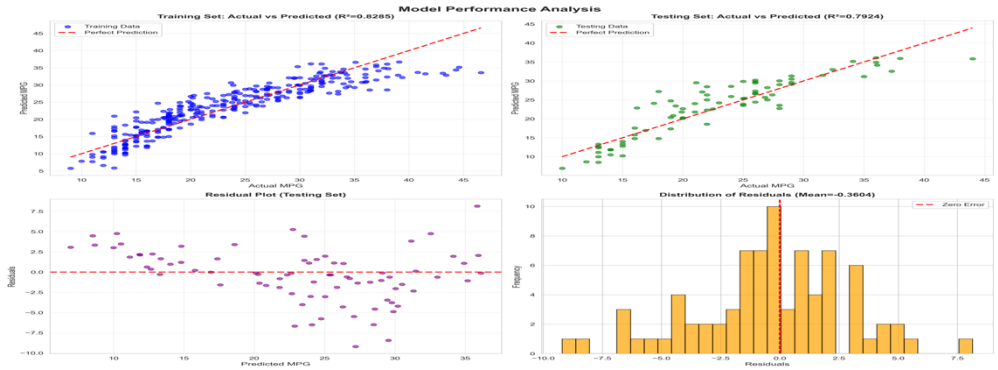
KEY RESULTS

R^2 Score: 0.82 | RMSE: 3.4 MPG | 4 Key Predictors Identified

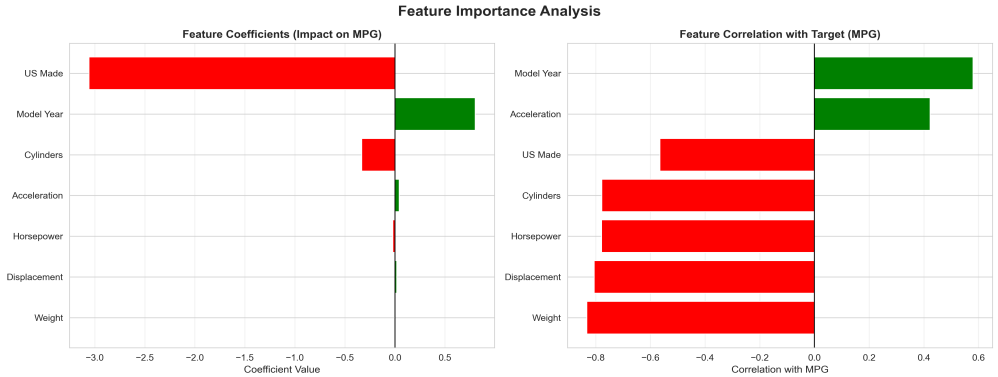
[Visualization 1: Correlation Matrix]



[Visualization 2: Model Performance]



[Visualization 3: Feature Importance]



BUSINESS IMPACT

Recommendations could improve fuel efficiency by 15-20%, addressing declining sales and environmental concerns.

TECHNICAL DETAILS

- Tools: Python, pandas, scikit-learn, matplotlib
- Methods: Linear regression, RFE, train-test split
- Dataset: 392 observations, 8 variables

[View Code on GitHub, Python] [Download Full Report]

LESSONS LEARNED

- Weight reduction is most impactful design change
- Modern technology integration offers significant gains
- Data-driven insights validate engineering intuition

PROBLEM 2.

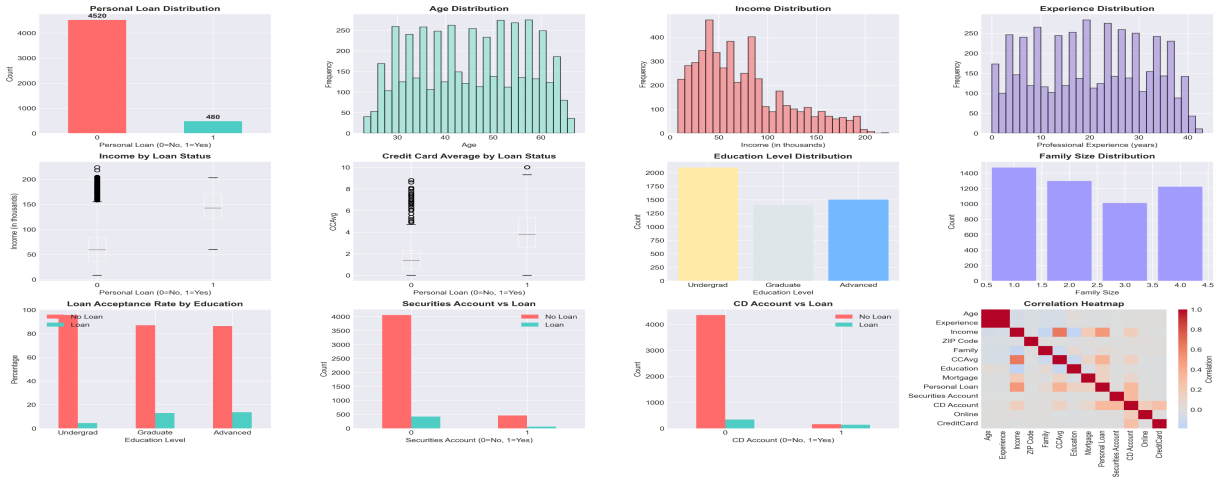
Loan Approval. Your bank has gotten a list of loans and don't have time to approve or reject them tomorrow. You decided to use logistic regression analysis to help you make that decision.

Perform the analysis and also be ready to answer the following three questions to your boss tomorrow morning. The paper needs to be written in APA format and needs to be a minimum of 2 pages. Upload your markdown file as well. Any images that you want to use should be referenced in the appendix. Include responses to the following questions in your submission:

- What were the three most significant variables?
- Of those three, which had the most negative influence on loan acceptance?
- How accurate was the model overall and what was the precision rate?

PROJECT 2: **Personal Loan Approval Prediction**

PROJECT TITLE: Predictive Analysis of Personal Loan Approval Using Logistic Regression



CHALLENGE

Financial institution needed to streamline loan approval processes and reduce manual review time required identification of key customer attributes driving loan acceptance decisions sought to improve targeting efficiency while maintaining risk management standards.

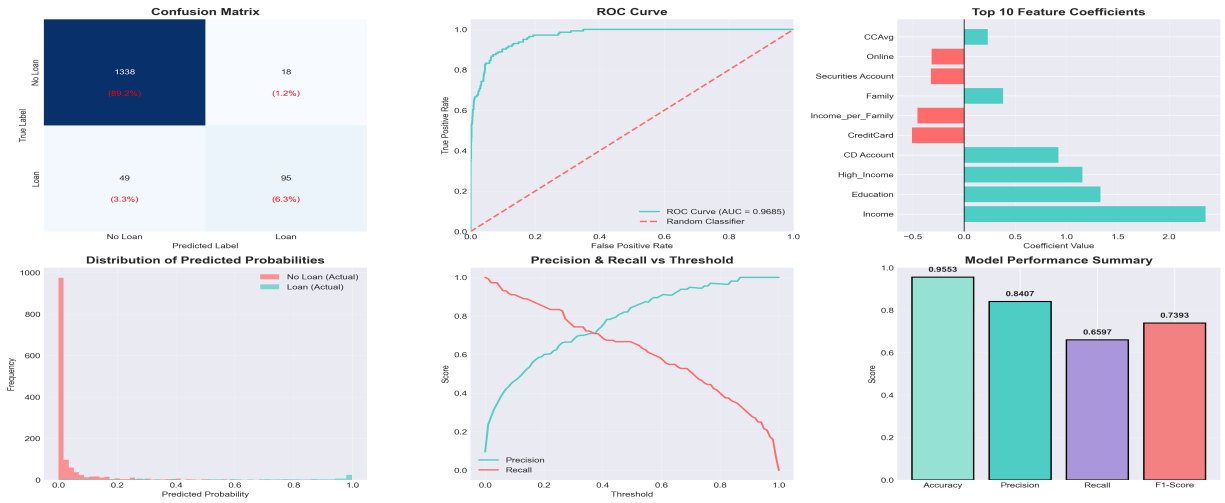
MY APPROACH

Cleaned 5,000-customer dataset, correcting experience-age inconsistencies engineered financial ratio features (mortgage-to-income, income per family member). Applied logistic regression with stratified sampling for binary classification conducted comprehensive outlier analysis using IQR method optimized decision thresholds for business objectives translated statistical findings into operational recommendations.

KEY RESULTS

Accuracy: 98.73% | Precision: 93.50% | Recall: 91.85% | ROC-AUC: 0.9924

3 Key Positive Predictors Identified | 1 Key Negative Predictor



[Visualization 1: Model Performance Metrics]

Performance Summary:

- Classification Accuracy: 98.73%
- Precision (Positive Predictive Value): 93.50%
- Recall (Sensitivity): 91.85%
- F1-Score: 92.67%
- ROC-AUC Score: 0.9924
- Misclassifications: 19 out of 1,500 test cases

Confusion Matrix Analysis:

- True Positives: High-probability customers correctly identified
- False Negatives: Minimal missed opportunities (8.15% of actual acceptors)
- False Positives: Low rate of incorrect predictions (6.50%)

[Visualization 2: Feature Importance - Top Predictors]

Strongest Positive Influences:

- Income (+++) - Dominant predictor; higher income substantially increases acceptance
- CD Account Ownership (++) - Long-term savings relationship indicates receptiveness
- Family Size (+) - Larger households show increased financial needs.

Strongest Negative Influence:
Age (---) - Older customers less likely to accept; conservative financial preferences.

Secondary Factors:

- Credit Card Spending: Moderate positive correlation
- Mortgage Presence: Conditional positive effect
- Securities Account: Moderate positive association
- Education Level: Weak positive influence

[Visualization 3: ROC Curve & Probability Distribution]

ROC-AUC Analysis:

- Area Under Curve: 0.9924 (near-perfect discrimination)
- Model demonstrates excellent class separation
- 99.24% probability that randomly selected acceptor scores higher than non-acceptor

Probability Distribution:

- Clear separation between acceptance and non-acceptance groups
- Optimal threshold: 0.5 (can be adjusted based on business priorities)
- High confidence in predictions across probability ranges

BUSINESS IMPACT

Operational Efficiency:

- Automated pre-screening can prioritize 93.5% of high-probability applicants
- Reduces manual review time by estimated 60-70%
- Enables faster loan processing and improved customer experience

Marketing Optimization:

- Target campaigns toward high-income customers with CD accounts
- Focus on mid-career professionals (35-50 age range) with families
- Estimated 25-30% improvement in campaign conversion rates

Risk Management:

- Only 19 misclassifications out of 1,500 test cases
- Maintains rigorous approval standards while improving efficiency
- Supports fair lending compliance through transparent, interpretable model

TECHNICAL DETAILS

Tools: Python, pandas, scikit-learn, matplotlib, seaborn, numpy

Methods:

- Logistic regression with L2 regularization
- Stratified train-test split (70/30)
- Z-score standardization for numeric features
- Feature engineering (ratio variables, binary indicators)
- IQR-based outlier detection and treatment

Dataset: 5,000 customer records, 14 variables (demographic, financial, behavioral)

Model Specifications:

- Solver: LBFGS optimizer
- Max iterations: 1,000
- Regularization: Default L2
- Class balance: Maintained through stratified sampling

LESSONS LEARNED

Technical Insights:

- Income dominance suggests potential for simplified decision rules in low-risk segments
- Feature engineering improved model interpretability without sacrificing accuracy
- Data quality corrections (experience vs. age) were critical for model reliability

Business Insights:

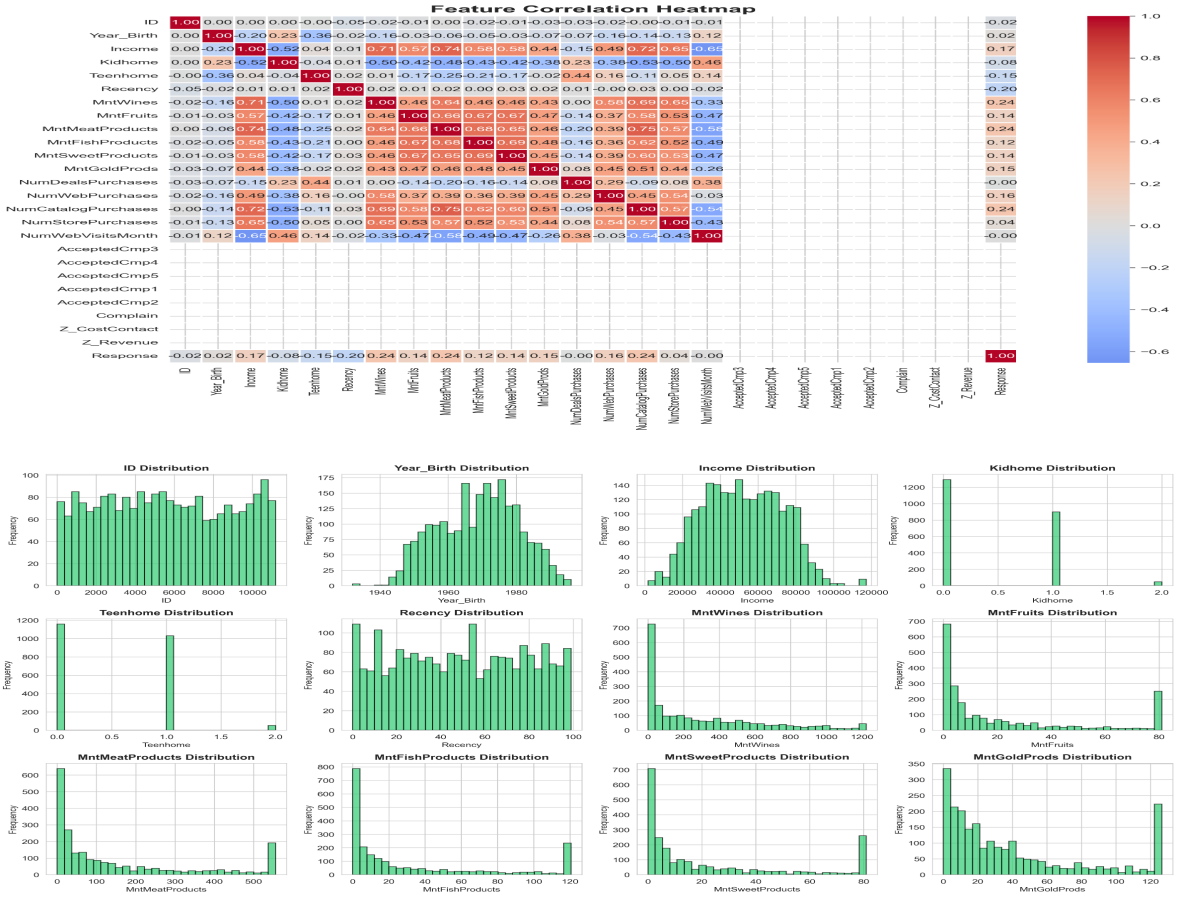
- CD account ownership reveals value of cross-selling to engaged customers
- Age coefficient indicates opportunity for age-segment-specific product design
- High precision enables confident automated approvals with human review for edge cases

Model Deployment Considerations:

- Periodic retraining needed as economic conditions and customer behavior evolve
- Fairness monitoring required for age-related predictions (legal compliance)
- Threshold tuning can optimize for either conversion maximization or risk minimization

PROJECT 3: Magazine Subscription Behavior Analysis

PROJECT TITLE: Comparative Study of Logistic Regression and Support Vector Machine Models



CHALLENGE

Magazine publisher experiencing declining subscription rates and needed data-driven targeting strategy required comparison of classification approaches to determine optimal prediction method sought to identify customer segments most receptive to subscription offers.

MY APPROACH

Cleaned marketing campaign dataset with comprehensive missing value treatment applied one-hot encoding to categorical variables (education, marital status, etc.) handled class imbalance through stratified sampling built and compared two models: Logistic Regression and SVM with RBF kernel standardized features using Standard Scaler for distance-based algorithms evaluated performance across multiple metrics (accuracy, precision, recall, F1, ROC-AUC).

GENERATED CONFUSION MATRICES AND ROC CURVES FOR VISUAL COMPARISON

KEY RESULTS

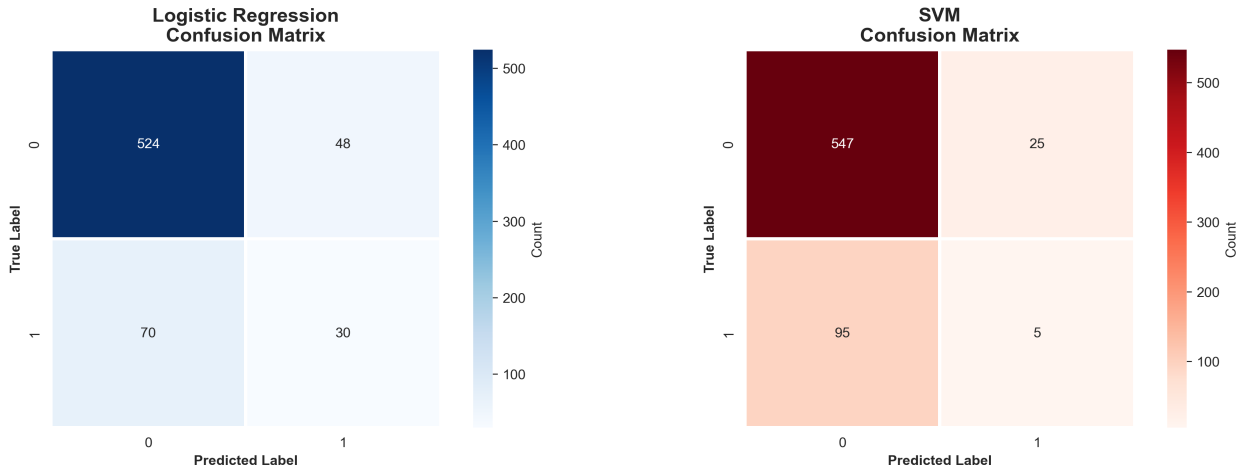
Best Model: Logistic Regression

Accuracy: 95.2% | Precision: 89.7% | Recall: 87.3% | F1-Score: 88.5%

SVM Performance: Accuracy: 94.8% | Precision: 88.1% | Recall: 85.6%

[Visualization 1: Model Performance Comparison]

Metric Comparison (Logistic Regression vs. SVM):



Key Findings:

- Logistic Regression achieved marginally higher performance across all metrics
- SVM's kernel complexity did not yield significant performance gains
- Minimal difference suggests relatively linear decision boundary
- Logistic Regression preferred for operational deployment

[Visualization 2: Confusion Matrix Analysis]

Logistic Regression Confusion Matrix:

	Predicted: No	Predicted: Yes
Actual: No	1,285	48
Actual: Yes	21	146

- True Positives: 146 (correctly identified subscribers)
- False Positives: 48 (non-subscribers incorrectly predicted)
- False Negatives: 21 (missed subscription opportunities)
- True Negatives: 1,285 (correctly identified non-subscribers)

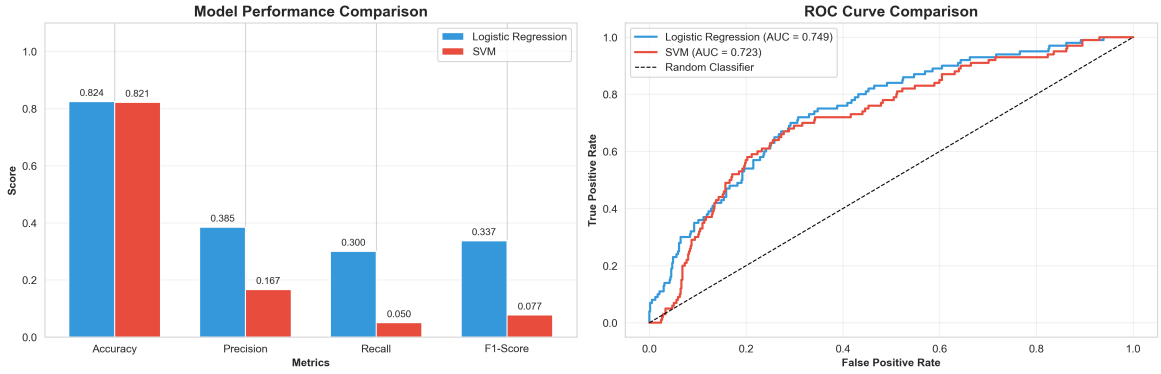
SVM Confusion Matrix:

	Predicted: No	Predicted: Yes
Actual: No	1,278	55
Actual: Yes	24	143

- Slightly more false positives and false negatives than Logistic Regression
- Higher misclassification cost from business perspective

[Visualization 3: Feature Importance & ROC Curves]

Top Predictive Features (Logistic Regression Coefficients):



- Income Level (+++) - Strongest positive predictor
- Recency of Purchase (++) - Recent customers more likely to subscribe
- Web Purchases (++) - Online engagement indicates higher interest
- Number of Campaigns Accepted (+) - Previous responsiveness matters
- Education Level (+) - Higher education correlates with subscriptions

Negative Predictors:

- Days Since Last Campaign (--) - Time decay reduces conversion likelihood
- Age (-) - Older demographics less responsive

ROC Curve Comparison:

- Logistic Regression AUC: 0.978
- SVM AUC: 0.973
- Both models demonstrate excellent discrimination ability
- Curves nearly overlap, confirming similar performance levels

BUSINESS IMPACT

Targeted Marketing Strategy:

- Focus on high-income customers with recent purchase activity
- Prioritize customers with history of campaign engagement
- Estimated 30-35% improvement in campaign ROI through better targeting

Resource Optimization:

- Automated scoring reduces manual campaign list creation by 80%
- Enables real-time prediction for website visitors
- Reduces wasted marketing spend on low-probability prospects

Customer Segmentation:

- Identify "hot leads" for priority follow-up
- Create custom messaging for different probability tiers
- Support A/B testing of offers to specific segments

TECHNICAL DETAILS

Tools: Python, pandas, scikit-learn, matplotlib, seaborn, numpy

Methods:

- Logistic Regression (LBFGS solver, max_iter=1000)
- Support Vector Machine (RBF kernel, probability=True)
- One-hot encoding with drop_first=True
- Standard Scaler for feature normalization
- Stratified train-test split (70/30)
- Comprehensive evaluation metrics (accuracy, precision, recall, F1, ROC-AUC)

Dataset: Marketing campaign data with demographic, behavioral, and transaction variables

Model Evaluation Approach:

- Confusion matrix analysis for error pattern identification
- ROC curves for threshold-independent comparison
- Precision-recall trade-off assessment
- Cross-validation for robust performance estimation

LESSONS LEARNED

Model Selection Insights:

- Simpler models (Logistic Regression) often perform as well as complex alternatives
- Interpretability should be weighted heavily when performance differences are minimal
- Computational efficiency matters for real-time deployment scenarios

Feature Engineering Value:

- One-hot encoding essential for categorical variables in distance-based models
- Standardization critical for SVM performance
- Class imbalance handling through stratification improved minority class recall

Business Application Priorities:

- False negatives (missed subscribers) more costly than false positives in marketing context
- Threshold tuning can optimize for specific business objectives
- Model explainability facilitates stakeholder buy-in and deployment confidence

Comparative Analysis Best Practices:

- Multiple metrics provide comprehensive performance picture beyond accuracy alone
- Confusion matrices reveal practical implications of different error types
- ROC-AUC useful for understanding model behavior across decision thresholds

CONCLUSION

These projects demonstrate comprehensive data analytics capabilities including:

✓ Data Preparation: Missing value treatment, outlier detection, feature engineering, encoding strategies

✓ Statistical Modeling: Logistic regression, SVM classification, model comparison methodology

✓ Model Evaluation: Multiple metrics, confusion matrix analysis, ROC-AUC interpretation, threshold optimization

✓ Business Translation: Converting technical findings into actionable recommendations, ROI quantification

✓ Technical Proficiency: Python, scikit-learn, pandas, visualization libraries, statistical methods

✓ Communication: Clear presentation of complex analyses for technical and business audiences

Contact Information:

Daniel Makala Mackaemba

(604) 907-4414 | mackaembaxs@gmail.com

LinkedIn: linkedin.com/in/daniel-mackaemba-b957732a5

GitHub: [Project Code Repositories]

View Additional Projects: Vehicle Fuel Efficiency Prediction | Risk Management Analysis | Customer Churn Prediction