TITLE: CREDIT RISK CLASSIFICATION MODEL FOR LOAN APPROVAL

Subtitle: Leveraging Data to Optimize Loan Approvals and Reduce Risks

Presented by: Kartik Ukhalkar

BUSINESS PROBLEM

- **Objective:** Develop a classification model to differentiate loan applicants into approved and rejected categories.
- **Key Questions:**
  - How to leverage trade-level information from Credit Bureaus?
  - Which application or payment behavior factors significantly influence borrower behavior?
  - How can these factors inform decision-making strategies?

DATASET OVERVIEW

- **Datasets Used:**
  - Application Data: Information about applicants (e.g., income, credit amount).
  - Bureau Data: Trade-level credit bureau records.
- **Initial Observations:**
  - Application Data: 307,511 rows, 43 features after preprocessing.
  - Bureau Data: Aggregated to applicant level with key statistics like active loans, overdue amounts.

DATA PREPROCESSING AND FEATURE ENGINEERING

Steps Taken:

- Dropped features with >40% missing values.
- Imputed missing values with mean/median (numerical) or mode (categorical).
- Aggregated bureau data to applicant level using statistical metrics (e.g., mean, max, count).
- Removed features with very low correlation to the target variable.

EXPLORATORY DATA ANALYSIS (EDA)

- **Insights:**
  - **Negative DAYS variables:** Indicate days before the loan application date (e.g., DAYS_CREDIT).
  - Significant predictors:
    - External Sources (EXT_SOURCE_2 and EXT_SOURCE_3) have strong correlation with TARGET.
    - Number of credit bureau inquiries (AMT_REQ_CREDIT_BUREAU_YEAR).
  - Class imbalance observed in TARGET: ~92% approved, ~8% rejected.

MODELS DEVELOPED

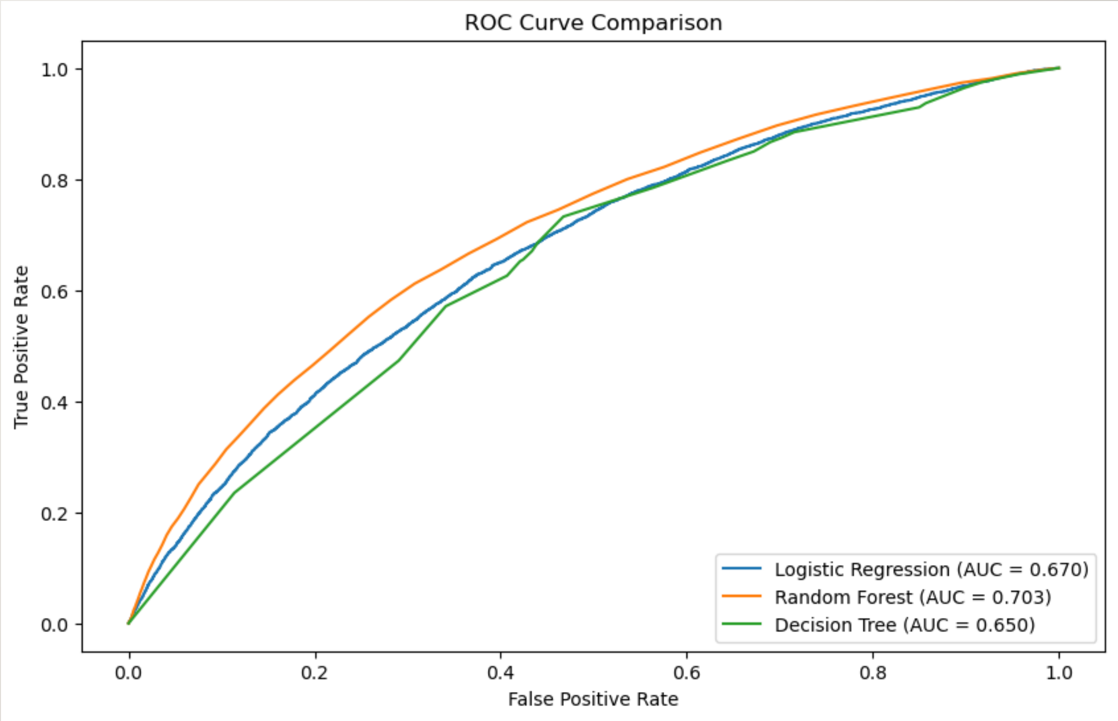
- **Classification Models:**
  - Logistic Regression
  - Decision Tree
  - Random Forest
- **Class Imbalance Handling:**
  - Oversampling using SMOTE (Synthetic Minority Oversampling Technique).

MODEL PERFORMANCE

- Evaluation Metric: ROC-AUC Score
- Results: Logistic Regression: 0.670
- Random Forest: 0.703 (Best Model)
- Decision Tree: 0.658
- Reason for Selection: Random Forest

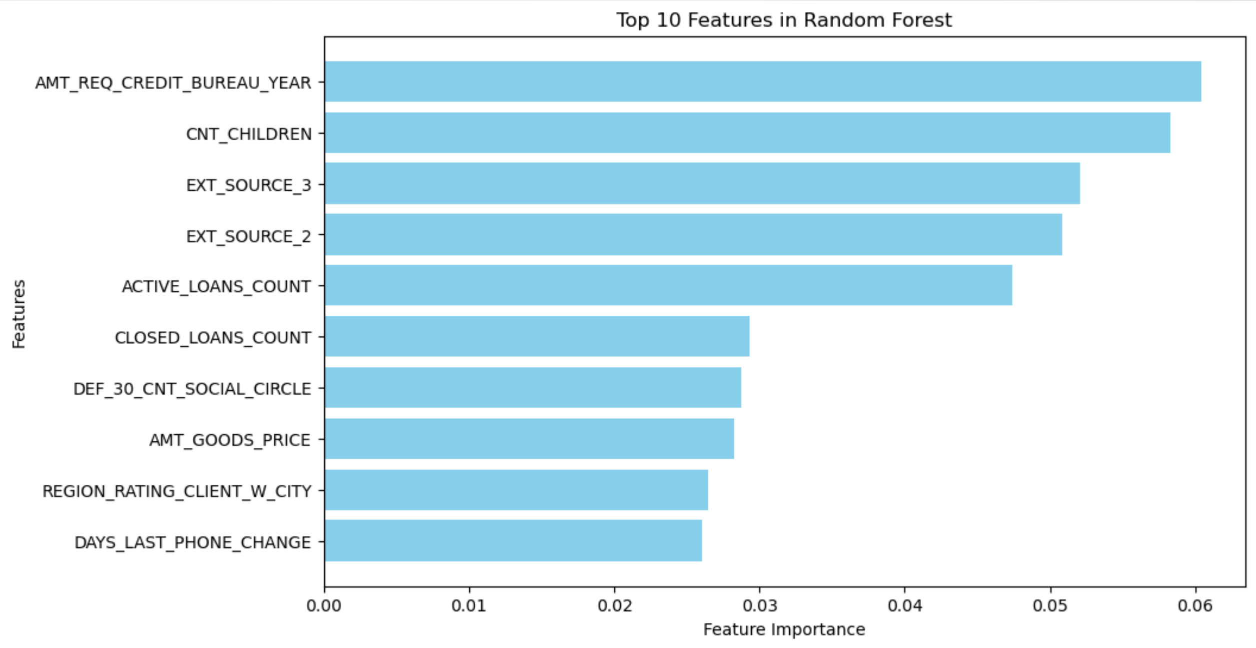
offers the best trade-off between

performance and interpretability.



TOP INFLUENTIAL FEATURES

- **Key Features Identified (Top 10):**
  - AMT_REQ_CREDIT_BUREAU_YEAR
  - CNT_CHILDREN
  - EXT_SOURCE_3
  - EXT_SOURCE_2
  - ACTIVE_LOANS_COUNT
  - CLOSED_LOANS_COUNT
  - DEF_30_CNT_SOCIAL_CIRCLE
  - AMT_GOODS_PRICE
  - REGION_RATING_CLIENT_W_CITY
  - DAYS_LAST_PHONE_CHANGE



- **Business Insight:** These features highlight critical behavioral and financial traits of applicants.

BUSINESS STRATEGIES

- **High-Risk Applicants:**
  - Stricter lending terms (e.g., higher interest rates, smaller loan amounts).
  - Increased scrutiny on payment history and income stability.
- **Low-Risk Applicants:**
  - Proactive targeting with larger loan amounts or longer terms.
  - Incentives like lower interest rates.
- **Threshold-Based Decisioning:**
  - Approve: Probability > 0.8
  - Manual Review: 0.5 – 0.8
  - Reject: Probability < 0.5

VISUALIZATION OF INSIGHTS

- **Feature Importance:** Bar chart of top features.
- **ROC Curve:** Model's ability to differentiate between classes.
- **Risk Score Distribution:** Histogram showing segmentation of applicants into low, medium, and high risk.

BUSINESS IMPLICATIONS

- Enhanced Risk Management: Identify high-risk applicants early to mitigate loan defaults.
- Optimized Loan Approvals: Increase approval rates for low-risk applicants while maintaining profitability.
- Data-Driven Decisioning: Use model insights to guide policy adjustments and product offerings.

NEXT STEPS

1. **Deploy Model:** Integrate into the bank's decision-making process.
2. **Monitor Performance:** Continuously evaluate and retrain the model as needed.
3. **Refine Strategies:** Use insights to develop new financial products or adjust lending policies.
4. **Stakeholder Training:** Ensure teams understand how to interpret and act on model outputs.