CREDIT CARD APPROVAL PREDICTION

Team: Lainey Li, Chris Lin,
Amy Wang, Olivia Wang, JiaWei Zhang

By leveraging ML, banks can improve quality and efficiency of credit card approval processes

Business Problem

1. Challenge: Balancing between correct approvals and risk avoidance
2. The evolving financial landscape: Digital finance and unpredictability
3. Ensuring fairness, transparency, and compliance in decisions

Adding Business Value through Analytics

1. Emphasis on data science & ML's role in credit decisioning
2. Transforming traditional evaluation methods
3. Adapting to changing economic conditions & ensuring compliance

Use Scenario in Analytics

1. Harnessing diverse applicant data for deeper insights
2. Predictive analytics for informed credit decisions

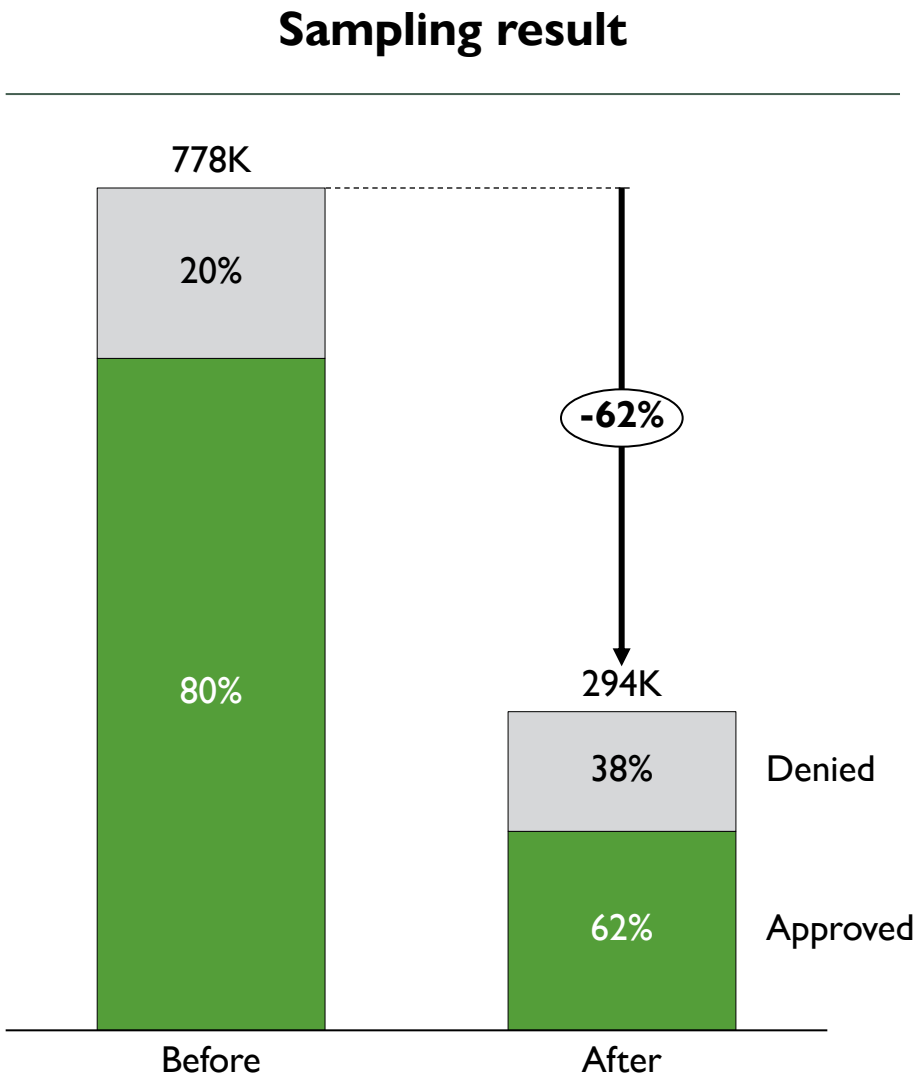
Good applicants are defined as those who pay off dues within 30 days, cut-off point that achieves balance between risk and profits

Existing target variable

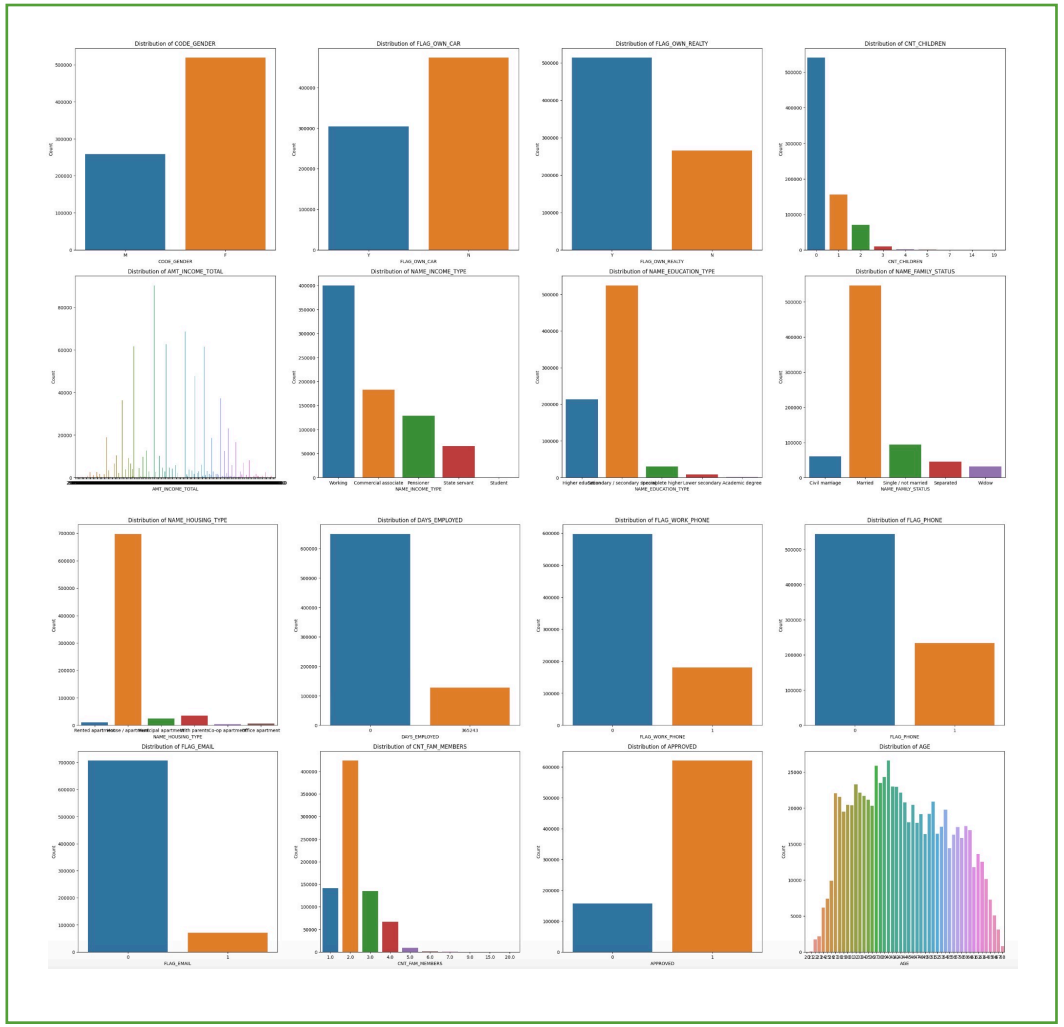
- C: paid off that month
- 0: 1-29 days past due

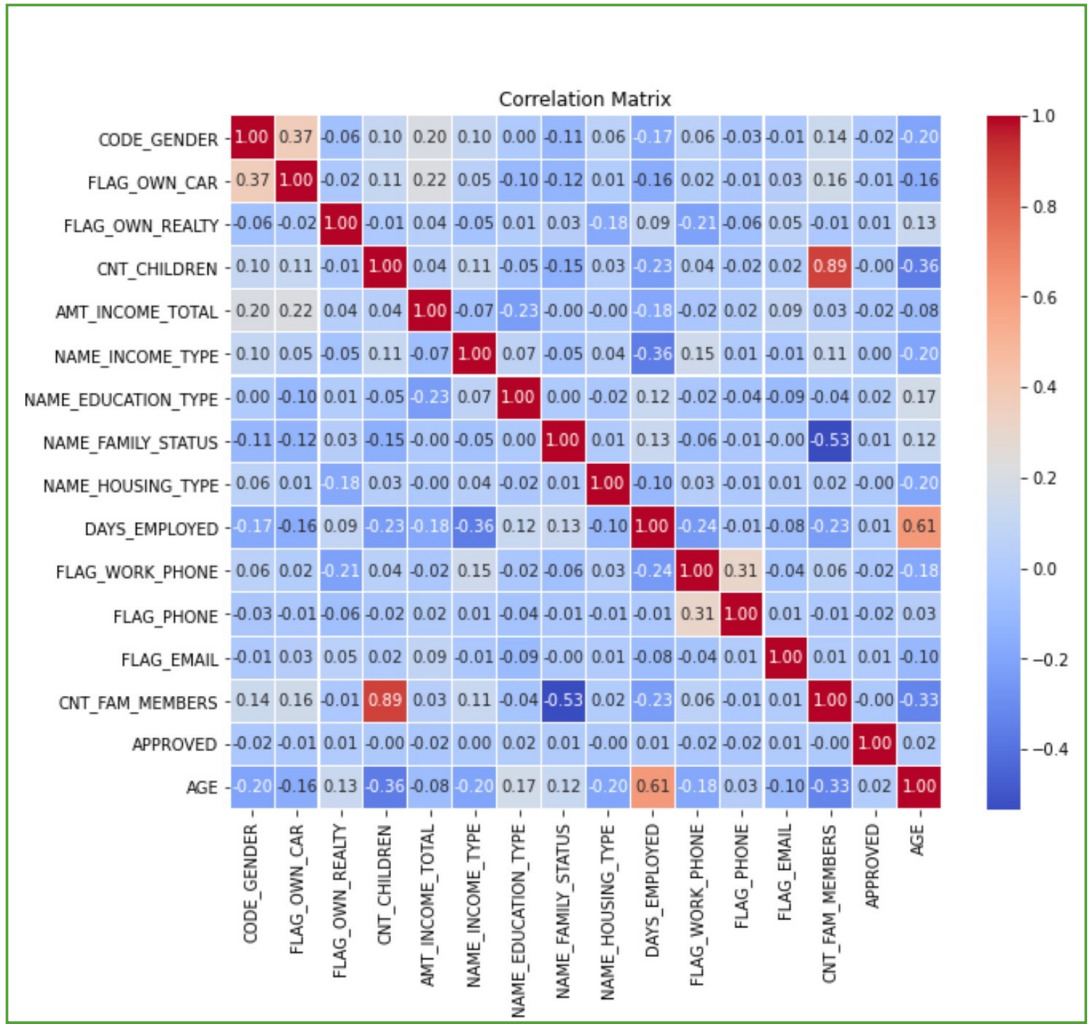
- 1: 30-59 days past due
- 2: 60-89 days overdue
- 3: 90-119 days overdue
- 4: 120-149 days overdue
- 5: Overdue or bad debts for more than 150 days
- X: No loan for the month

Undersampling imbalance data brings about 3 benefits



Exploratory data analysis

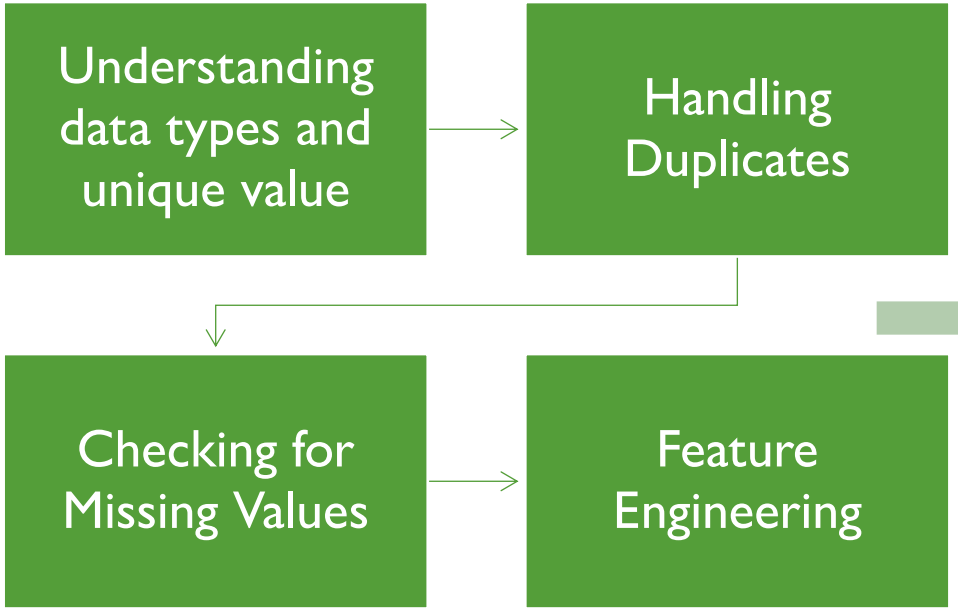




4 categories of variables can help determine whether to approve or reject credit card applicants

Category	Demographics	Economic status	Education/Job	Contact info
Variable				
Hypothesis				

4 steps of processing are performed to ensure data quality



- Duplicated IDs were dropped from the application record dataframe
- Unnecessary columns were dropped, including occupation type (missing values) and flag mobile (single value)
- The application record and credit record datasets were merged on the unique ID column to get a consolidated dataset
- A new feature, AGE, was engineered from the DAYS_BIRTH column
- Days employed were adjusted to represent only positive values, with unemployed days set to 0
- Encoding was done for categorical features to convert them into numerical values 0/1

Data Modeling

- **Type of Model:** Classification
- **Data Mining Algorithms**
  - Logistic Regression
  - K-Nearest Neighbors (KNN)
  - Decision Tree

F1 score with Cross Validation

Model	KNN	Logistic Regression	Decision Tree
F1 Score	0.77	0.42	0.83

Imbalanced Classes

F1 addresses skewed datasets where accuracy can mislead

Cost of Errors

False Positives: Potential losses from unpaid credit

False Negatives: Lost revenue opportunities

Balancing Precision & Recall

Ensures neither metric is sacrificed for the other

Comparative Evaluation

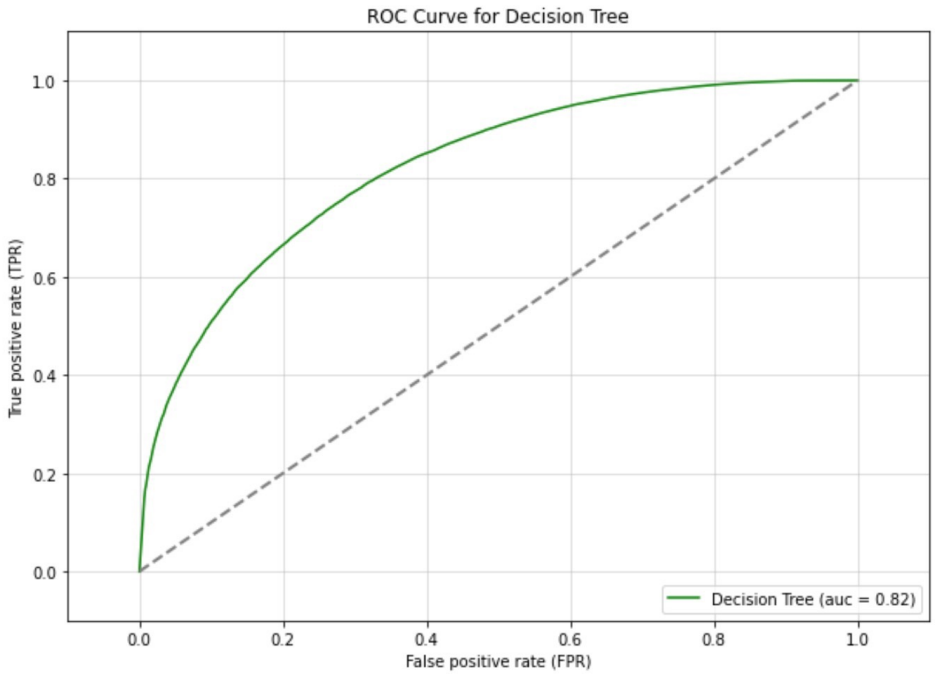
Single, comprehensive metric for model tweaking & comparison

Decision Tree outperformed

TUNING

Utilize a grid search for hyperparameter tuning for Decision Tree, with nested cross validation

- 5 folds in the inner and outer loop
- F-1 scoring metric
- Decision Tree Nested CV F1 score: 0.81 +/- 0.002
- An AUC of 0.82 suggests that the Decision Tree classifier is quite effective, but not perfect, in distinguishing between the two classes



KEY TAKEAWAY: ROC and AUC provide an overall assessment of a classifier's performance across various thresholds, while F1 score gives a single metric that balances precision and recall. They are complementary evaluation measures that can be used to assess the effectiveness of a ML model.

Potential issues need to be addressed before implementation

Issues to Consider

- Periodic re-training to maintain relevance
- Management of false positives and negatives

Ethical Considerations

- Avoidance of bias against specific applicant groups
- Upholding applicants' data privacy

Risks and Mitigation:

- False Positives: Review and adjust threshold values periodically
- Model Outdatedness: Regular retraining and validation with fresh data
- Data Breaches: Employ data encryption and robust access controls

The model can be integrated into the bank's loan application system to predict and flag potential high-risk applications in real-time