Home Credit Default Risk

Dan Powell, Matt Johnescu, Melissa Messervy

Overview

- Problem Statement
- Model Performance
- Limitations
- Future Considerations
- Business Implications

Overview

- Problem Statement
- Model Performance
- Limitations
- Future Considerations
- Business Implications

Problem Statement

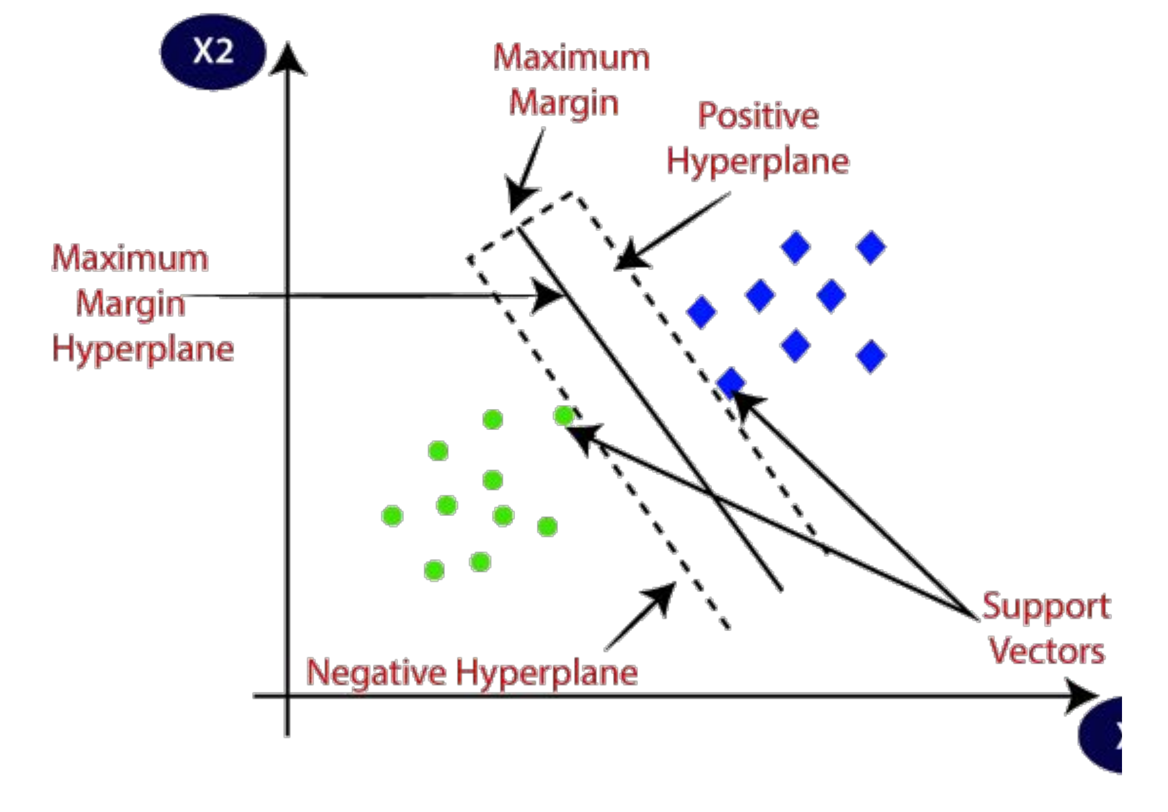
- **Home Credit's Mission:** Targeting underserved populations - People with insufficient or no credit history.
- **Problem**: Difficulty obtaining loans or exposure to untrustworthy lenders.
- **Solution**: Using alternative data to improve financial inclusion and provide a safe, positive loan experience.

Overview

- Problem Statement
- Model Performance
- Limitations
- Future Considerations
- Business Implications

Support Vector Machine (SVM)

- **Models Tested:**
  - **Linear Model** (without SMOTE)
  - **Linear Model** (with SMOTE)
  - **Radial Model** (with SMOTE)
- **Key Terms:**
  - **SMOTE**: Synthetic Minority Oversampling Technique
  - **Linear Model**: Divides classes linearly
  - **Radial Model**: Divides classes non-linearly
- **Performance Metrics:**
  - **Accuracy**: Frequency of correct predictions
  - **F1 Score**: Measure of model's ability to classify positive cases



SVM Modeling Process and Results

Model	Accuracy	F1 Score
Linear SVM (No Adjustments)	91.93%	N/A
Linear SVM (with Adjustments)	33.75%	0.1719
Radial SVM (with adjustments and optimization)	**63.96%**	**0.175**

Linear Regression

Low R-squared values: LM (0.06) and Rpart (0.02)

High RMSE & Relative Absolute Error

5-fold Cross Validation was performed with minimal change in RMSE and RAE

Model Performances

Model	Kaggle Score
SVM	0.58357
Linear Regression for Classification	0.73867
LightGBM Model	0.74085

LightGBM Model Overview

- Tree based based classification with gradient boosted learning
- Tree Construction
- Gradient Boosting
- Optimization and Prediction
- Low Memory Consumption

Model Implementation

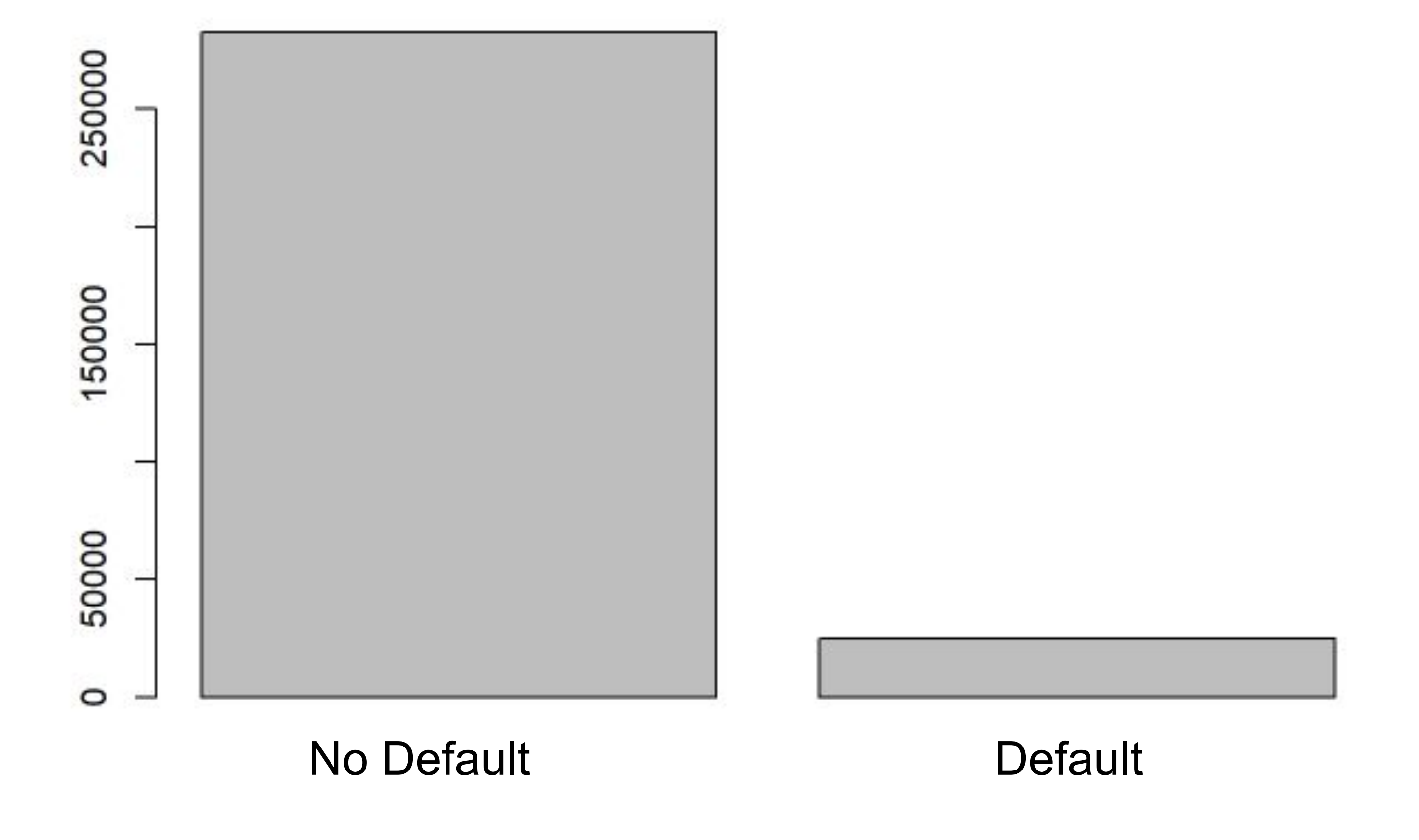
- Feature Engineering
- Cross Validation
- Early Stopping
- Class Imbalance

Overview

- Problem Statement
- Model Performance
- Limitations
- Future Considerations
- Business Implications

Target Variable Imbalance

Loan Default Distribution



Model Limitations

- Computational Limitations
- Data Quality
- Interpretability
- Data Updates
- Explore Additional Model types

Overview

Problem Statement

Model Performance

Limitations

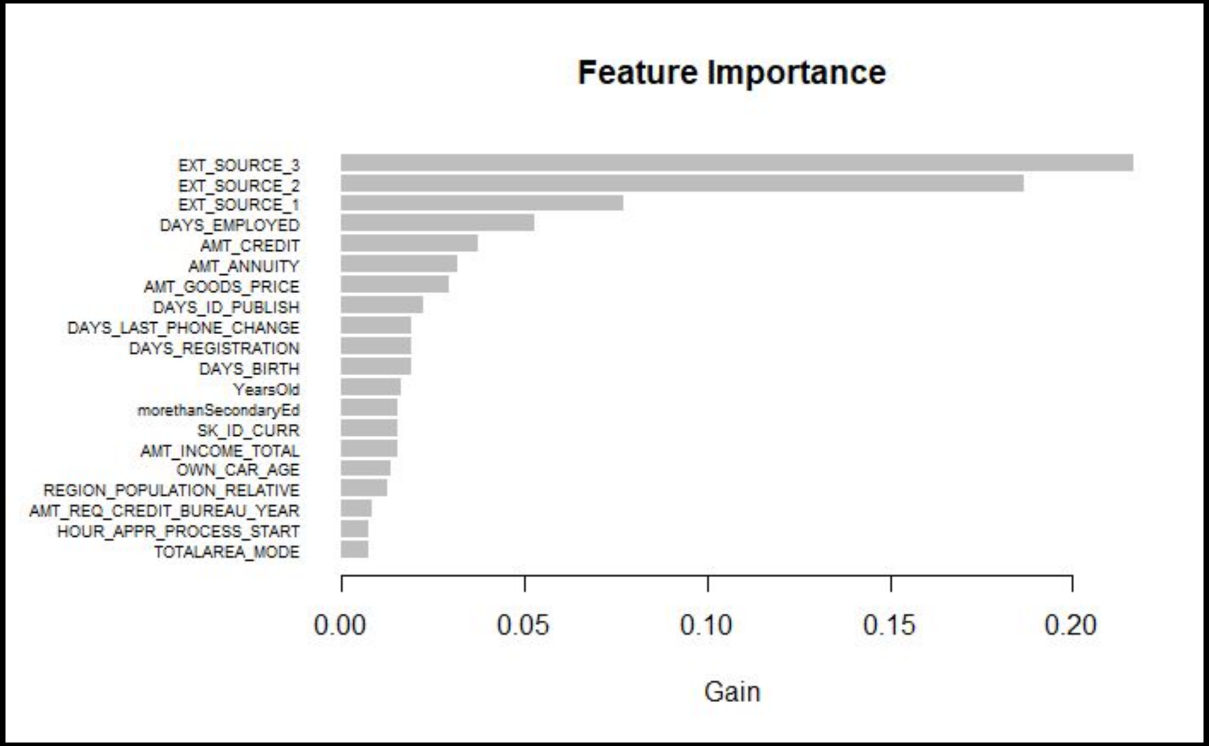
Future Considerations

Business Implications

Future Considerations & Implications

Real-time Data Cleaning: Analyze customer loan default probability

Auto Hyperparameter Tuning: Enhance model performance



Overview

- Problem Statement
- Model Performance
- Limitations
- Future Considerations
- Business Implications

Business Implications

KEY PREDICTORS TO IMPROVE LOAN DEFAULT DECISIONS

CUSTOMER DATA IN REAL TIME

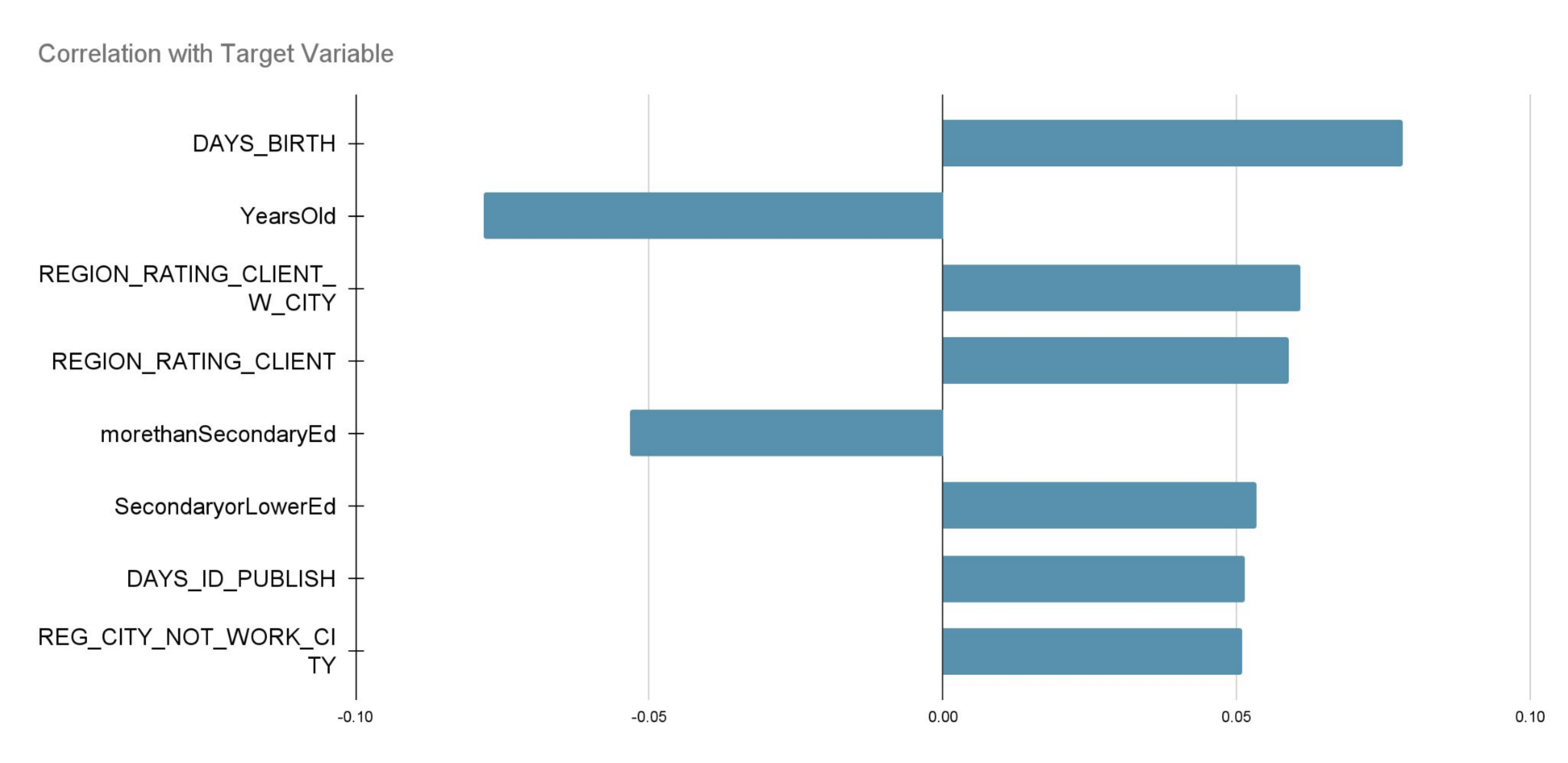
FINANCIAL INCLUSION

SMARTER LENDING PRACTICES

Questions?

Appendix

Variables of Interest



Variables of Interest

Categorical Variables of interest:

- Custom Variables:
  - IsMarried
  - **SecondaryEducation**
  - isCashLoan
- Demographic and financial information

Binary Variables of interest

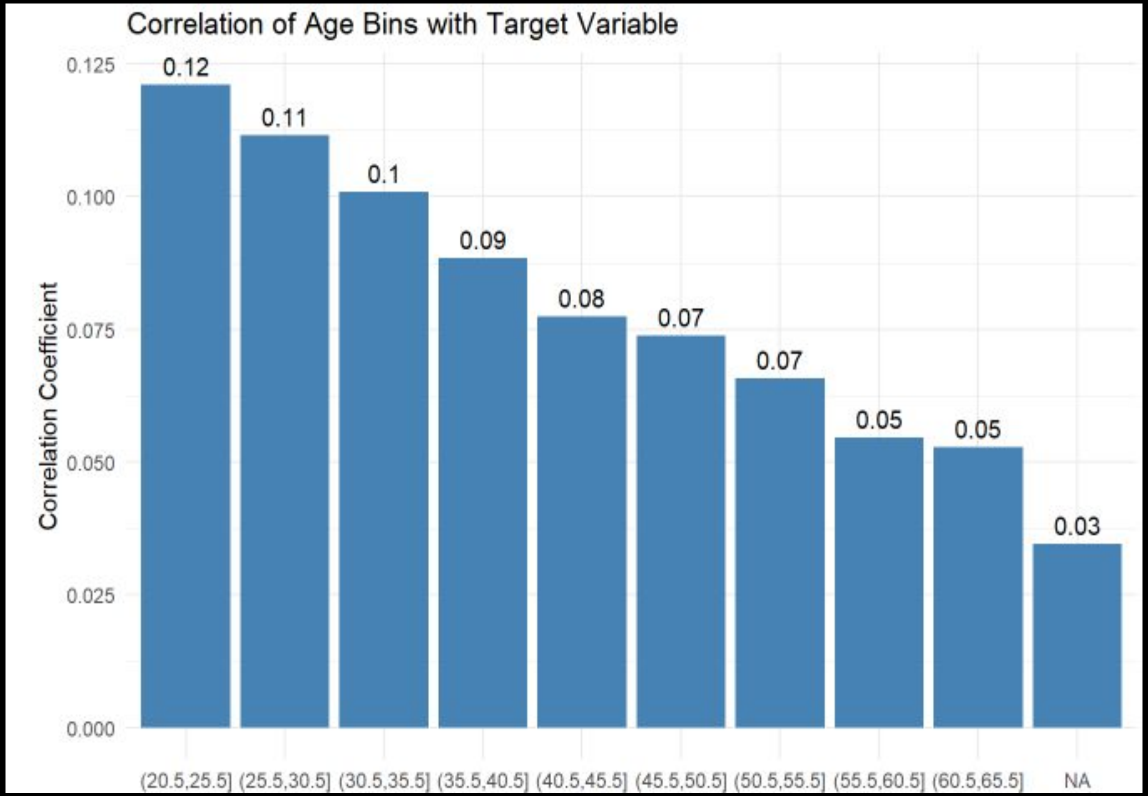
- Document flags, having a work phone, etc

Age Groups from DAYS_BIRTH

- Correlation with age and target

Numerical variables of interest

- Region rating
- Days employed



EDA – Pre-Modeling Data Cleaning

- Binary Columns
  - Created 4 new factor columns:
    - isMarried, isCashLoan, morethanSecondaryEd, DAY_EMPLOYED_ANOM
- Numerical Columns
  - Transform days into years for years old
- Factor Columns
  - Transform Years old into buckets by every 10 years
  - Replace missing values with a “missing” level

Pre-Modeling Transformations - SVM

- Extract top 20 features most correlated with the target
- Missing values
  - Median impute numerical values with NAs
  - Add “missing” category to age_group NAs
- One-Hot Encode Categorical Variables
- Scale Data around mean of 0 for numerical (non-binary) columns
- Randomly sampled 5000 observations from training data for performance

"TARGET"	"DAYS_BIRTH"
"YearsOld"	"REGION_RATING_CLIENT_W_CITY"
"REGION_RATING_CLIENT"	"morethanSecondaryEd"
"SecondaryorLowerEd"	"DAYS_ID_PUBLISH"
"REG_CITY_NOT_WORK_CITY"	"DAY_EMPLOYED_ANOM"
"FLAG_EMP_PHONE"	"REG_CITY_NOT_LIVE_CITY"
"FLAG_DOCUMENT_3"	"DAYS_REGISTRATION"
"REGION_POPULATION_RELATIVE"	"LIVE_CITY_NOT_WORK_CITY"
"isCashLoan"	"AMT_CREDIT"
"FLAG_DOCUMENT_6"	"FLAG_WORK_PHONE"
"HOUR_APPR_PROCESS_START"	"FLAG_PHONE"
"CNT_CHILDREN"	"isMarried"
"FLAG_DOCUMENT_16"	"FLAG_DOCUMENT_13"
"DAYS_LAST_PHONE_CHANGE"	"AMT_ANNUITY"
"AMT_GOODS_PRICE"	"age_group_20_30"
"age_group_30_40"	"age_group_40_50"
"age_group_50_60"	"age_group.age_group_missing"

Linear Model without SMOTE - SVM

- F1 Score: N/A
- Accuracy: 91.93%
- All predictions for No Class
    - Caused by class imbalance
- Model is essentially useless

		Reference	
Prediction		N	Y
	N	56537	4965
	Y	0	0

Confusion Matrix for Linear Model

Linear Model With SMOTE

- Applying SMOTE:
  - Creates synthetic observations of the minority class
- Trained the same model but on SMOTE Data
- Model output:
  - F1 Score: 0.1719
  - Accuracy: 0.3375
- Less accurate but better at catching actual positives

SMOTE

N	Y
4599	401

N	Y
4010	4411

	Reference	
Prediction	N	Y
N	16525	736
Y	40012	4229

Confusion Matrix for Linear Model with SMOTE

Radial Model with SMOTE (Best Performing)

- Now use Radial Model with SMOTE, weights and optimized threshold
  - Can adjust for non-linear classification tasks
- Model Results on Test Data
  - Accuracy: 0.6396
  - F1 Score: 0.175
- Much better at picking true positives than prior models, but still has room for improvement for both false and true positives.
- Best Performing SVM Model

		Reference	
Prediction		N	Y
	N	36984	2614
	Y	19553	2351

Confusion Matrix for Radial Model