Home Credit Project

Predicting Consumer Credit Defaults

Druce Vertes
Metis Data Science Bootcamp
October 31, 2018

The Problem

Company

Home Credit is an international consumer lender, founded in the Czech Republic, that uses alternative data to model and issue credit cards and consumer loans.

Context

Lending to someone who pays on time is profitable.

Lending to someone who doesn't pay it back is unprofitable.

8.1% of approved loans in this data set defaulted.

Problem

Predict likelihood of default using

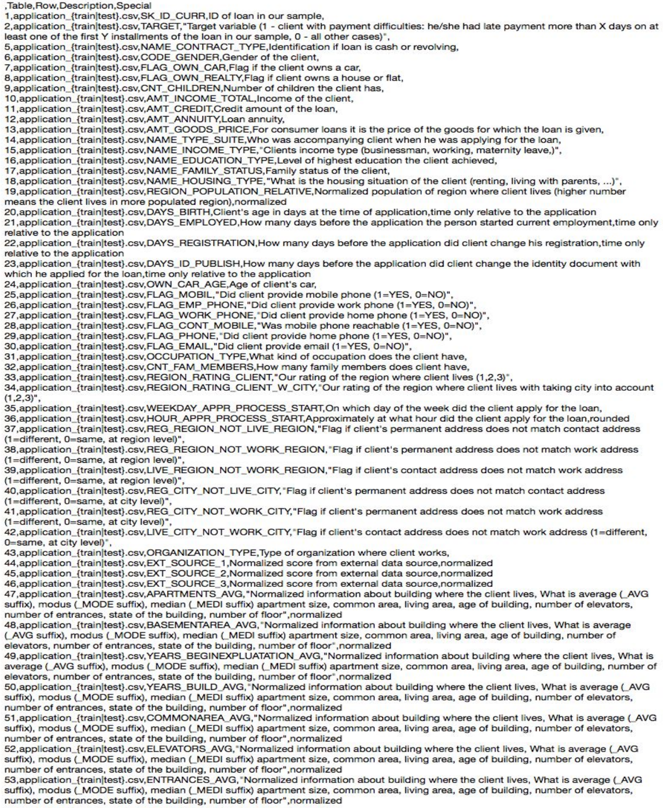
- Application data
- Previous applications
- Credit reports
- Payment history

Target Variable:

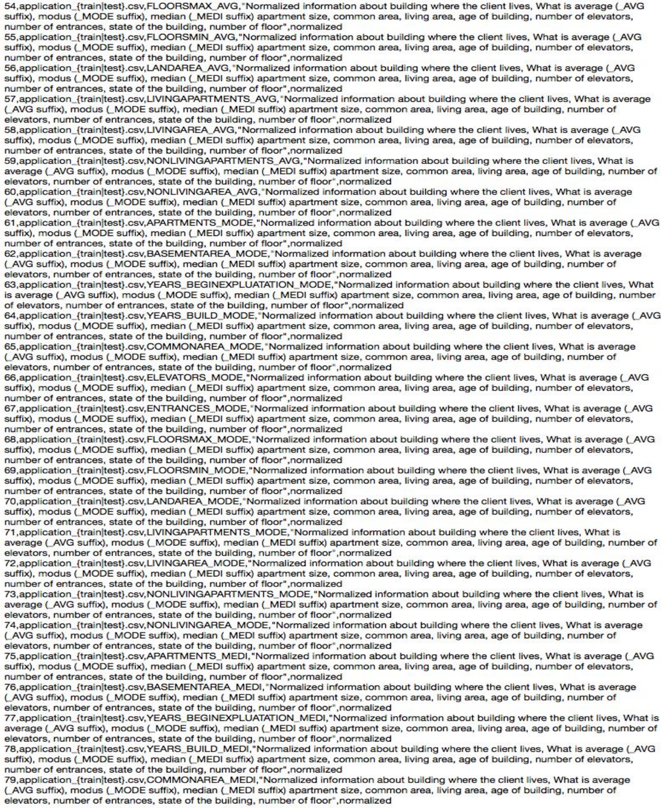
1 - Client with payment difficulties: he/she had late payment more than X days on at least one of the first Y installments of the loan in our sample

0 - all other cases

Features (221)



```
,Table,Row,Description,Special
1,application_{train|test}.csv,SK_ID_CURR,ID of loan in our sample,
2,application_{train|test}.csv,TARGET,"Target variable (1 - client with payment difficulties: he/she had late payment more than X days on at
least one of the first Y installments of the loan in our sample, 0 - all other cases)",
5,application_{train|test}.csv,NAME_CONTRACT_TYPE,Identification if loan is cash or revolving,
6,application_{train|test}.csv,CODE_GENDER,Gender of the client,
7,application_{train|test}.csv,FLAG_OWN_CAR,Flag if the client owns a car,
8,application_{train|test}.csv,FLAG_OWN_REALTY,Flag if client owns a house or flat,
9,application_{train|test}.csv,CNT_CHILDREN,Number of children the client has,
10,application_{train|test}.csv,AMT_INCOME_TOTAL,Income of the client,
11,application_{train|test}.csv,AMT_CREDIT,Credit amount of the loan,
12,application_{train|test}.csv,AMT_ANNUITY,Loan annuity,
13,application_{train|test}.csv,AMT_GOODS_PRICE,For consumer loans it is the price of the goods for which the loan is given,
14,application_{train|test}.csv,NAME_TYPE_SUITE,Who was accompanying client when he was applying for the loan,
15,application_{train|test}.csv,NAME_INCOME_TYPE,"Clients income type (businessman, working, maternity leave,)",
16,application_{train|test}.csv,NAME_EDUCATION_TYPE,Level of highest education the client achieved,
17,application_{train|test}.csv,NAME_FAMILY_STATUS,Family status of the client,
18,application_{train|test}.csv,NAME_HOUSING_TYPE,"What is the housing situation of the client (renting, living with parents, ...)",
19,application_{train|test}.csv,REGION_POPULATION_RELATIVE,Normalized population of region where client lives (higher number
means the client lives in more populated region),normalized
20,application_{train|test}.csv,DAYS_BIRTH,Client's age in days at the time of application,time only relative to the application
21,application_{train|test}.csv,DAYS_EMPLOYED,How many days before the application the person started current employment,time only
relative to the application
22,application_{train|test}.csv,DAYS_REGISTRATION,How many days before the application did client change his registration,time only
relative to the application
23,application_{train|test}.csv,DAYS_ID_PUBLISH,How many days before the application did client change the identity document with
which he applied for the loan,time only relative to the application
24,application_{train|test}.csv,OWN_CAR_AGE,Age of client's car,
25,application_{train|test}.csv,FLAG_MOBIL,"Did client provide mobile phone (1=YES, 0=NO)",
26,application_{train|test}.csv,FLAG_EMP_PHONE,"Did client provide work phone (1=YES, 0=NO)",
27,application_{train|test}.csv,FLAG_WORK_PHONE,"Did client provide home phone (1=YES, 0=NO)",
28,application_{train|test}.csv,FLAG_CONT_MOBILE,"Was mobile phone reachable (1=YES, 0=NO)",
29,application_{train|test}.csv,FLAG_PHONE,"Did client provide home phone (1=YES, 0=NO)",
30,application_{train|test}.csv,FLAG_EMAIL,"Did client provide email (1=YES, 0=NO)",
31,application_{train|test}.csv,OCCUPATION_TYPE,What kind of occupation does the client have,
32,application_{train|test}.csv,CNT_FAM_MEMBERS,How many family members does client have,
33,application_{train|test}.csv,REGION_RATING_CLIENT,"Our rating of the region where client lives (1,2,3)",
34,application_{train|test}.csv,REGION_RATING_CLIENT_W_CITY,"Our rating of the region where client lives with taking city into account
(1,2,3)",
35,application_{train|test}.csv,WEEKDAY_APPR_PROCESS_START,On which day of the week did the client apply for the loan,
36,application_{train|test}.csv,HOUR_APPR_PROCESS_START,Approximately at what hour did the client apply for the loan,rounded
37,application_{train|test}.csv,REG_REGION_NOT_LIVE_REGION,"Flag if client's permanent address does not match contact address
(1=different, 0=same, at region level)",
38,application_{train|test}.csv,REG_REGION_NOT_WORK_REGION,"Flag if client's permanent address does not match work address
(1=different, 0=same, at region level)",
39,application_{train|test}.csv,LIVE_REGION_NOT_WORK_REGION,"Flag if client's contact address does not match work address
(1=different, 0=same, at region level)",
40,application_{train|test}.csv,REG_CITY_NOT_LIVE_CITY,"Flag if client's permanent address does not match contact address
(1=different, 0=same, at city level)",
41,application_{train|test}.csv,REG_CITY_NOT_WORK_CITY,"Flag if client's permanent address does not match work address
(1=different, 0=same, at city level)",
42,application_{train|test}.csv,LIVE_CITY_NOT_WORK_CITY,"Flag if client's contact address does not match work address (1=different,
0=same, at city level)",
43,application_{train|test}.csv,ORGANIZATION_TYPE,Type of organization where client works,
44,application_{train|test}.csv,EXT_SOURCE_1,Normalized score from external data source,normalized
45,application_{train|test}.csv,EXT_SOURCE_2,Normalized score from external data source,normalized
46,application_{train|test}.csv,EXT_SOURCE_3,Normalized score from external data source,normalized
47,application_{train|test}.csv,APARTMENTS_AVG,"Normalized information about building where the client lives, What is average (_AVG
suffix), modus (_MODE suffix), median (_MEDI suffix) apartment size, common area, living area, age of building, number of elevators,
number of entrances, state of the building, number of floor",normalized
48,application_{train|test}.csv,BASEMENTAREA_AVG,"Normalized information about building where the client lives, What is average
(_AVG suffix), modus (_MODE suffix), median (_MEDI suffix) apartment size, common area, living area, age of building, number of
elevators, number of entrances, state of the building, number of floor",normalized
49,application_{train|test}.csv,YEARS_BEGINEXPLUATATION_AVG,"Normalized information about building where the client lives, What is
average (_AVG suffix), modus (_MODE suffix), median (_MEDI suffix) apartment size, common area, living area, age of building, number of
elevators, number of entrances, state of the building, number of floor",normalized
50,application_{train|test}.csv,YEARS_BUILD_AVG,"Normalized information about building where the client lives, What is average (_AVG
suffix), modus (_MODE suffix), median (_MEDI suffix) apartment size, common area, living area, age of building, number of elevators,
number of entrances, state of the building, number of floor",normalized
51,application_{train|test}.csv,COMMONAREA_AVG,"Normalized information about building where the client lives, What is average (_AVG
suffix), modus (_MODE suffix), median (_MEDI suffix) apartment size, common area, living area, age of building, number of elevators,
number of entrances, state of the building, number of floor",normalized
52,application_{train|test}.csv,ELEVATORS_AVG,"Normalized information about building where the client lives, What is average (_AVG
suffix), modus (_MODE suffix), median (_MEDI suffix) apartment size, common area, living area, age of building, number of elevators,
number of entrances, state of the building, number of floor",normalized
53,application_{train|test}.csv,ENTRANCES_AVG,"Normalized information about building where the client lives, What is average (_AVG
suffix), modus (_MODE suffix), median (_MEDI suffix) apartment size, common area, living area, age of building, number of elevators,
number of entrances, state of the building, number of floor",normalized
```



```
54,application_{train|test}.csv,FLOORSMAX_AVG,"Normalized information about building where the client lives, What is average (_AVG
suffix), modus (_MODE suffix), median (_MEDI suffix) apartment size, common area, living area, age of building, number of elevators,
number of entrances, state of the building, number of floor",normalized
55,application_{train|test}.csv,FLOORSMIN_AVG,"Normalized information about building where the client lives, What is average (_AVG
suffix), modus (_MODE suffix), median (_MEDI suffix) apartment size, common area, living area, age of building, number of elevators,
number of entrances, state of the building, number of floor",normalized
56,application_{train|test}.csv,LANDAREA_AVG,"Normalized information about building where the client lives, What is average (_AVG
suffix), modus (_MODE suffix), median (_MEDI suffix) apartment size, common area, living area, age of building, number of elevators,
number of entrances, state of the building, number of floor",normalized
57,application_{train|test}.csv,LIVINGAPARTMENTS_AVG,"Normalized information about building where the client lives, What is average
(_AVG suffix), modus (_MODE suffix), median (_MEDI suffix) apartment size, common area, living area, age of building, number of
elevators, number of entrances, state of the building, number of floor",normalized
58,application_{train|test}.csv,LIVINGAREA_AVG,"Normalized information about building where the client lives, What is average (_AVG
suffix), modus (_MODE suffix), median (_MEDI suffix) apartment size, common area, living area, age of building, number of elevators,
number of entrances, state of the building, number of floor",normalized
59,application_{train|test}.csv,NONLIVINGAPARTMENTS_AVG,"Normalized information about building where the client lives, What is
average (_AVG suffix), modus (_MODE suffix), median (_MEDI suffix) apartment size, common area, living area, age of building, number of
elevators, number of entrances, state of the building, number of floor",normalized
60,application_{train|test}.csv,NONLIVINGAREA_AVG,"Normalized information about building where the client lives, What is average
(_AVG suffix), modus (_MODE suffix), median (_MEDI suffix) apartment size, common area, living area, age of building, number of
elevators, number of entrances, state of the building, number of floor",normalized
61,application_{train|test}.csv,APARTMENTS_MODE,"Normalized information about building where the client lives, What is average (_AVG
suffix), modus (_MODE suffix), median (_MEDI suffix) apartment size, common area, living area, age of building, number of elevators,
number of entrances, state of the building, number of floor",normalized
62,application_{train|test}.csv,BASEMENTAREA_MODE,"Normalized information about building where the client lives, What is average
(_AVG suffix), modus (_MODE suffix), median (_MEDI suffix) apartment size, common area, living area, age of building, number of
elevators, number of entrances, state of the building, number of floor",normalized
63,application_{train|test}.csv,YEARS_BEGINEXPLUATATION_MODE,"Normalized information about building where the client lives, What
is average (_AVG suffix), modus (_MODE suffix), median (_MEDI suffix) apartment size, common area, living area, age of building, number
of elevators, number of entrances, state of the building, number of floor",normalized
64,application_{train|test}.csv,YEARS_BUILD_MODE,"Normalized information about building where the client lives, What is average (_AVG
suffix), modus (_MODE suffix), median (_MEDI suffix) apartment size, common area, living area, age of building, number of elevators,
number of entrances, state of the building, number of floor",normalized
65,application_{train|test}.csv,COMMONAREA_MODE,"Normalized information about building where the client lives, What is average
(_AVG suffix), modus (_MODE suffix), median (_MEDI suffix) apartment size, common area, living area, age of building, number of
elevators, number of entrances, state of the building, number of floor",normalized
66,application_{train|test}.csv,ELEVATORS_MODE,"Normalized information about building where the client lives, What is average (_AVG
suffix), modus (_MODE suffix), median (_MEDI suffix) apartment size, common area, living area, age of building, number of elevators,
number of entrances, state of the building, number of floor",normalized
67,application_{train|test}.csv,ENTRANCES_MODE,"Normalized information about building where the client lives, What is average (_AVG
suffix), modus (_MODE suffix), median (_MEDI suffix) apartment size, common area, living area, age of building, number of elevators,
number of entrances, state of the building, number of floor",normalized
68,application_{train|test}.csv,FLOORSMAX_MODE,"Normalized information about building where the client lives, What is average (_AVG
suffix), modus (_MODE suffix), median (_MEDI suffix) apartment size, common area, living area, age of building, number of elevators,
number of entrances, state of the building, number of floor",normalized
69,application_{train|test}.csv,FLOORSMIN_MODE,"Normalized information about building where the client lives, What is average (_AVG
suffix), modus (_MODE suffix), median (_MEDI suffix) apartment size, common area, living area, age of building, number of elevators,
number of entrances, state of the building, number of floor",normalized
70,application_{train|test}.csv,LANDAREA_MODE,"Normalized information about building where the client lives, What is average (_AVG
suffix), modus (_MODE suffix), median (_MEDI suffix) apartment size, common area, living area, age of building, number of elevators,
number of entrances, state of the building, number of floor",normalized
71,application_{train|test}.csv,LIVINGAPARTMENTS_MODE,"Normalized information about building where the client lives, What is
average (_AVG suffix), modus (_MODE suffix), median (_MEDI suffix) apartment size, common area, living area, age of building, number of
elevators, number of entrances, state of the building, number of floor",normalized
72,application_{train|test}.csv,LIVINGAREA_MODE,"Normalized information about building where the client lives, What is average (_AVG
suffix), modus (_MODE suffix), median (_MEDI suffix) apartment size, common area, living area, age of building, number of elevators,
number of entrances, state of the building, number of floor",normalized
73,application_{train|test}.csv,NONLIVINGAPARTMENTS_MODE,"Normalized information about building where the client lives, What is
average (_AVG suffix), modus (_MODE suffix), median (_MEDI suffix) apartment size, common area, living area, age of building, number of
elevators, number of entrances, state of the building, number of floor",normalized
74,application_{train|test}.csv,NONLIVINGAREA_MODE,"Normalized information about building where the client lives, What is average
(_AVG suffix), modus (_MODE suffix), median (_MEDI suffix) apartment size, common area, living area, age of building, number of
elevators, number of entrances, state of the building, number of floor",normalized
75,application_{train|test}.csv,APARTMENTS_MEDI,"Normalized information about building where the client lives, What is average (_AVG
suffix), modus (_MODE suffix), median (_MEDI suffix) apartment size, common area, living area, age of building, number of elevators,
number of entrances, state of the building, number of floor",normalized
76,application_{train|test}.csv,BASEMENTAREA_MEDI,"Normalized information about building where the client lives, What is average
(_AVG suffix), modus (_MODE suffix), median (_MEDI suffix) apartment size, common area, living area, age of building, number of
elevators, number of entrances, state of the building, number of floor",normalized
77,application_{train|test}.csv,YEARS_BEGINEXPLUATATION_MEDI,"Normalized information about building where the client lives, What is
average (_AVG suffix), modus (_MODE suffix), median (_MEDI suffix) apartment size, common area, living area, age of building, number of
elevators, number of entrances, state of the building, number of floor",normalized
78,application_{train|test}.csv,YEARS_BUILD_MEDI,"Normalized information about building where the client lives, What is average (_AVG
suffix), modus (_MODE suffix), median (_MEDI suffix) apartment size, common area, living area, age of building, number of elevators,
number of entrances, state of the building, number of floor",normalized
79,application_{train|test}.csv,COMMONAREA_MEDI,"Normalized information about building where the client lives, What is average
(_AVG suffix), modus (_MODE suffix), median (_MEDI suffix) apartment size, common area, living area, age of building, number of
elevators, number of entrances, state of the building, number of floor",normalized
```

Feature Types

- Numerical: Income, credit score, payment amount

- Categorical: Income source, occupation

- Binary: M/F, own car

Baseline Model

- Application table data only
- No feature engineering: data as provided
- No model tuning, default hyperparameters
- Minimal scrubbing (a few weird outliers)
- Binarize Y/N, F/M, etc. -> 0,1
- Categorical -> one-hot dummies
- SimpleImputer to fill in numeric data

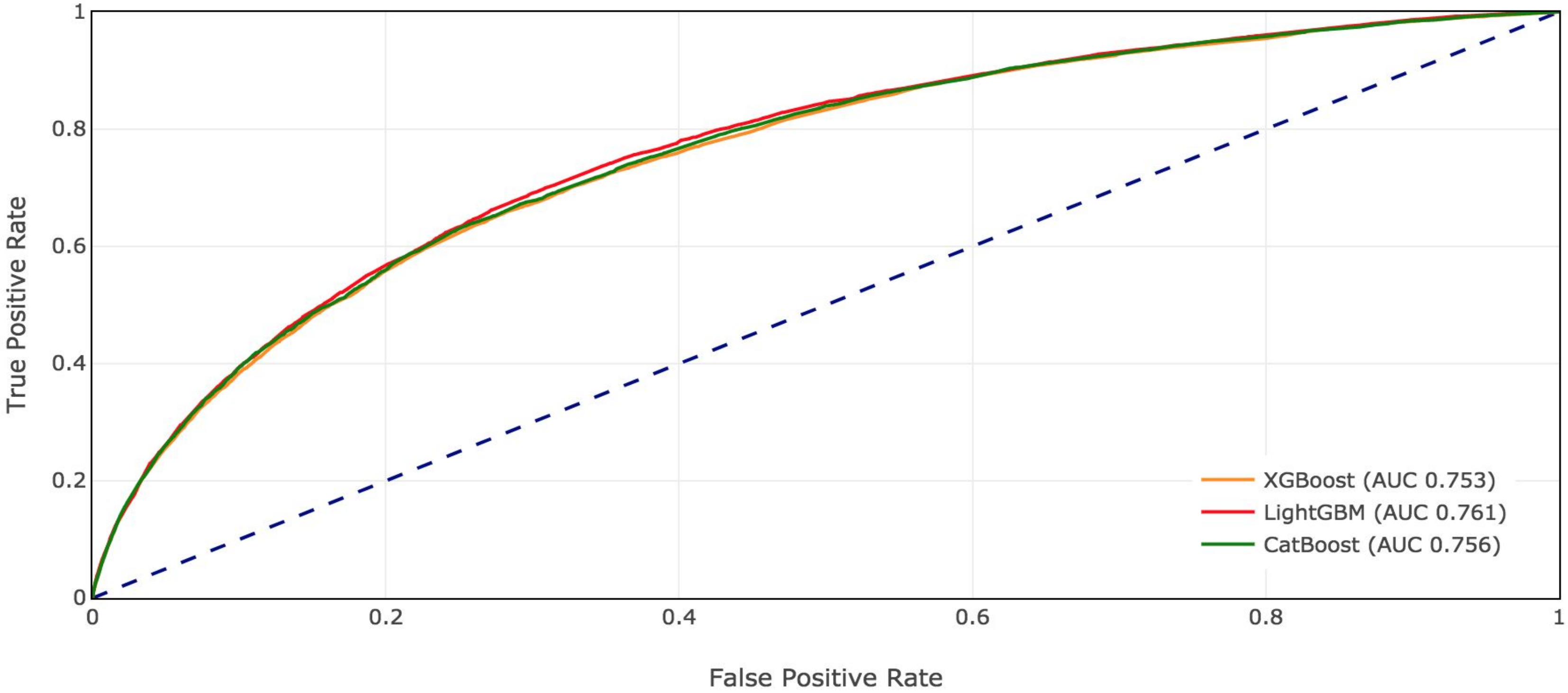
Problem

•**Baseline model**

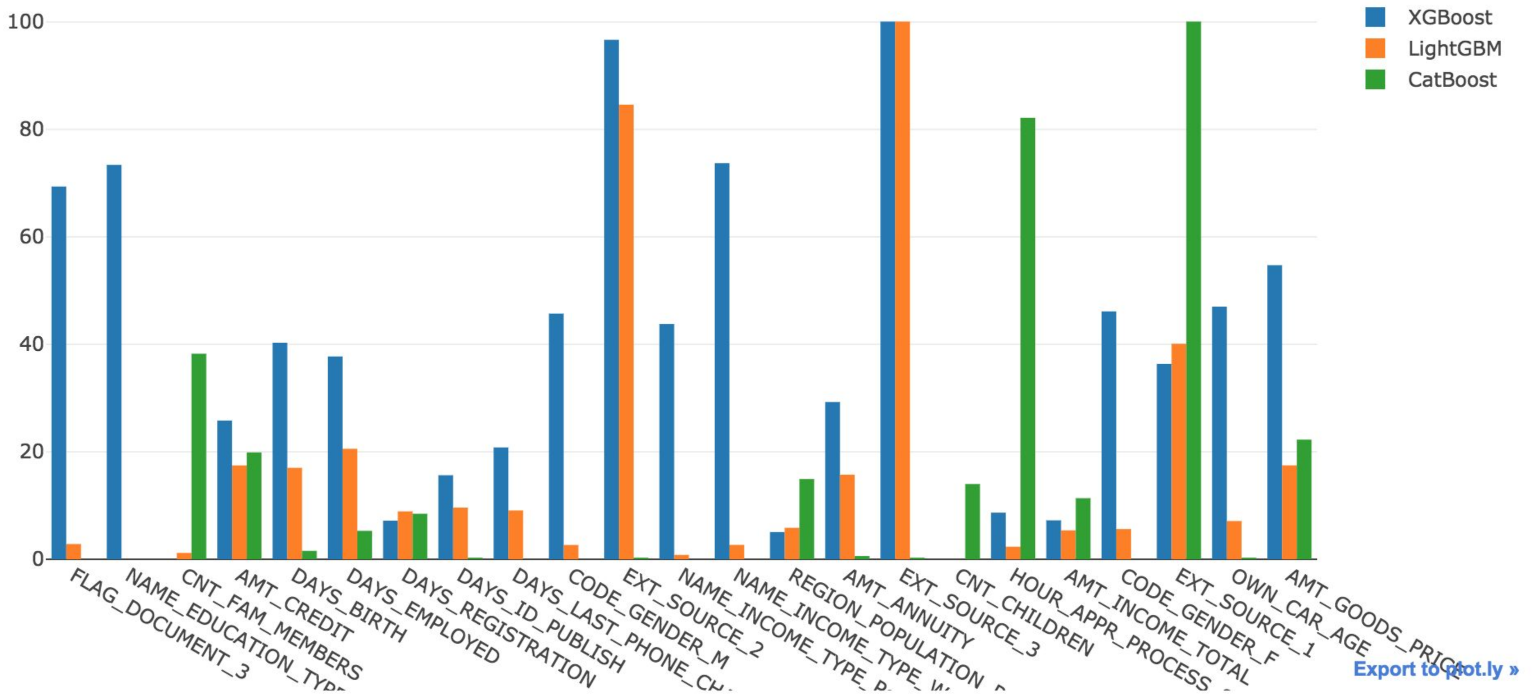
Improved model

Future improvements

Baseline Model ROC Curves Are Similar



Baseline Feature Importances Differ Widely



Baseline Model Metrics

Metric (Xval)	XGBoost	LightGBM	CatBoost
Accuracy	0.860	0.861	0.861
F1	0.305	0.311	0.311
AUC	0.753	0.760	0.757

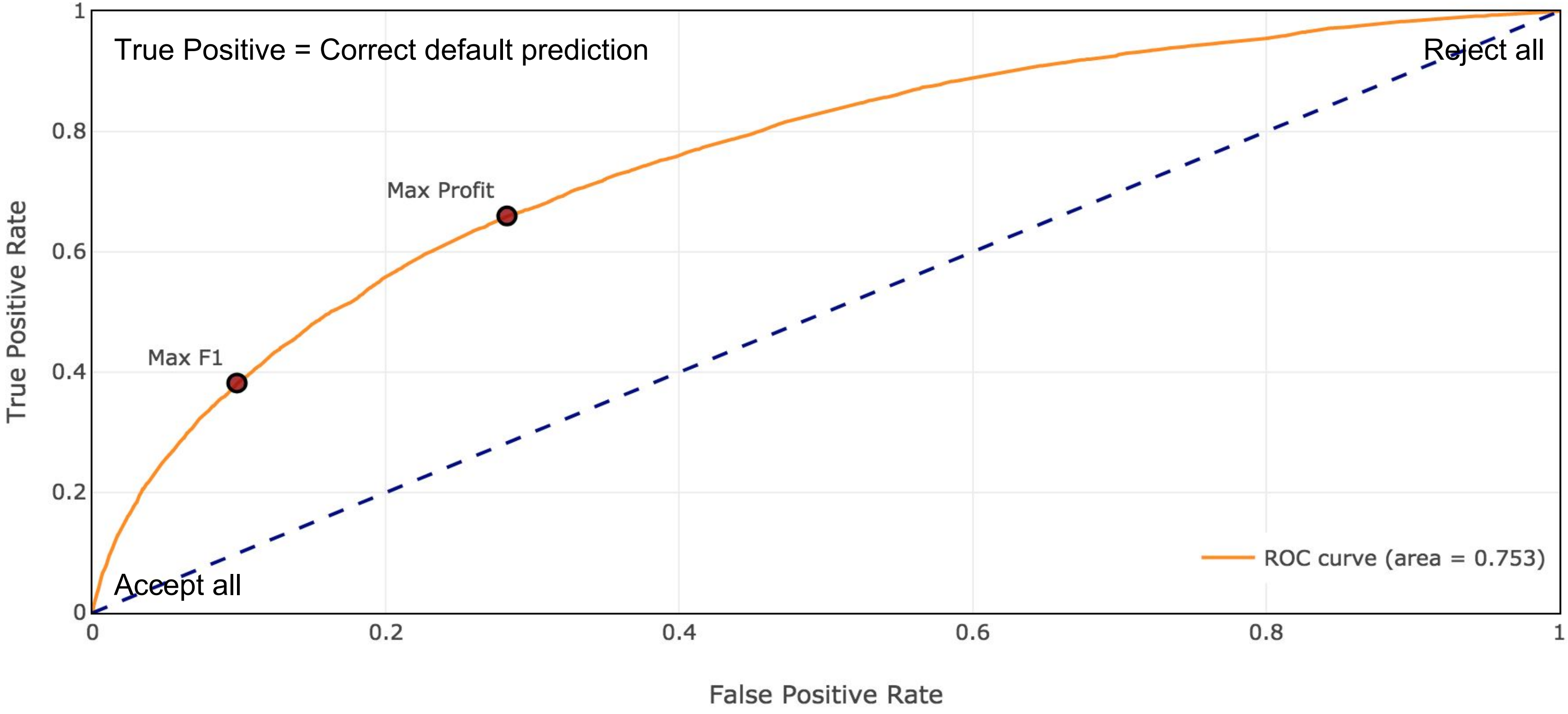
More Model Metrics - What Do We Care About?

- Metrics are abstractions
  - Deny all loans, F1=0.149 ; approve all, F1=0
  - But bank is (maybe?) better off making all loans
  - You can’t spend F1!
- We care about how much we make! Assign dollar values
  - Performing (Target=0) : $1,000 profit
  - Nonperforming (Target=1) : $11,387 loss
- Choose classification threshold to maximize total value (instead of F1)

More Model Metrics - Putting a Dollar Figure

Metric (Xval)	Base	XGBoost Best F1	XGBoost Best P/L
Accuracy	0.081	0.860	0.713
F1	0.149	0.305	0.271
Performing	56538	50971	40565
Nonperforming	4965	3068	1690
Value	0	+$16.0M	+$21.3M

ROC Curve - Max F1 vs. Max Profit



True Positive = Correct default prediction

Reject all

Accept all

Improved Model

Additional tables:

- Previous applications
- Credit bureau records
- Previous Home Credit accounts:
    - Credit card
    - Point-of-Sale (‘10 easy payments’ accounts)

Improved Model

Engineered features:

- Divide key amounts by reported income
- Aggregate historical tables
  - Count previous statuses
    - Applications, reasons for rejection
    - On time, late payments
  - Compute log1p where highly skewed (many counts of 0 late payments, some > 100)

Results - P/L Improvement

Metric (Xval)	Base	Baseline Best F1	Baseline Best P/L	Final Best P/L
Accuracy	0.081	0.860	0.713	0.684
F1	0.149	0.305	0.271	0.276
Performing	56538	50971	40565	38345
Nonperforming	4965	3068	1690	1260
Value	0	+$16.0M	+$21.3M	+24.0m

Results

- New features used
  - Past default status
  - Past accepted/refused (got better deal?)

- Kaggle AUC 0.78395
- This is not even the median of entries
- Contest winners are around 0.82

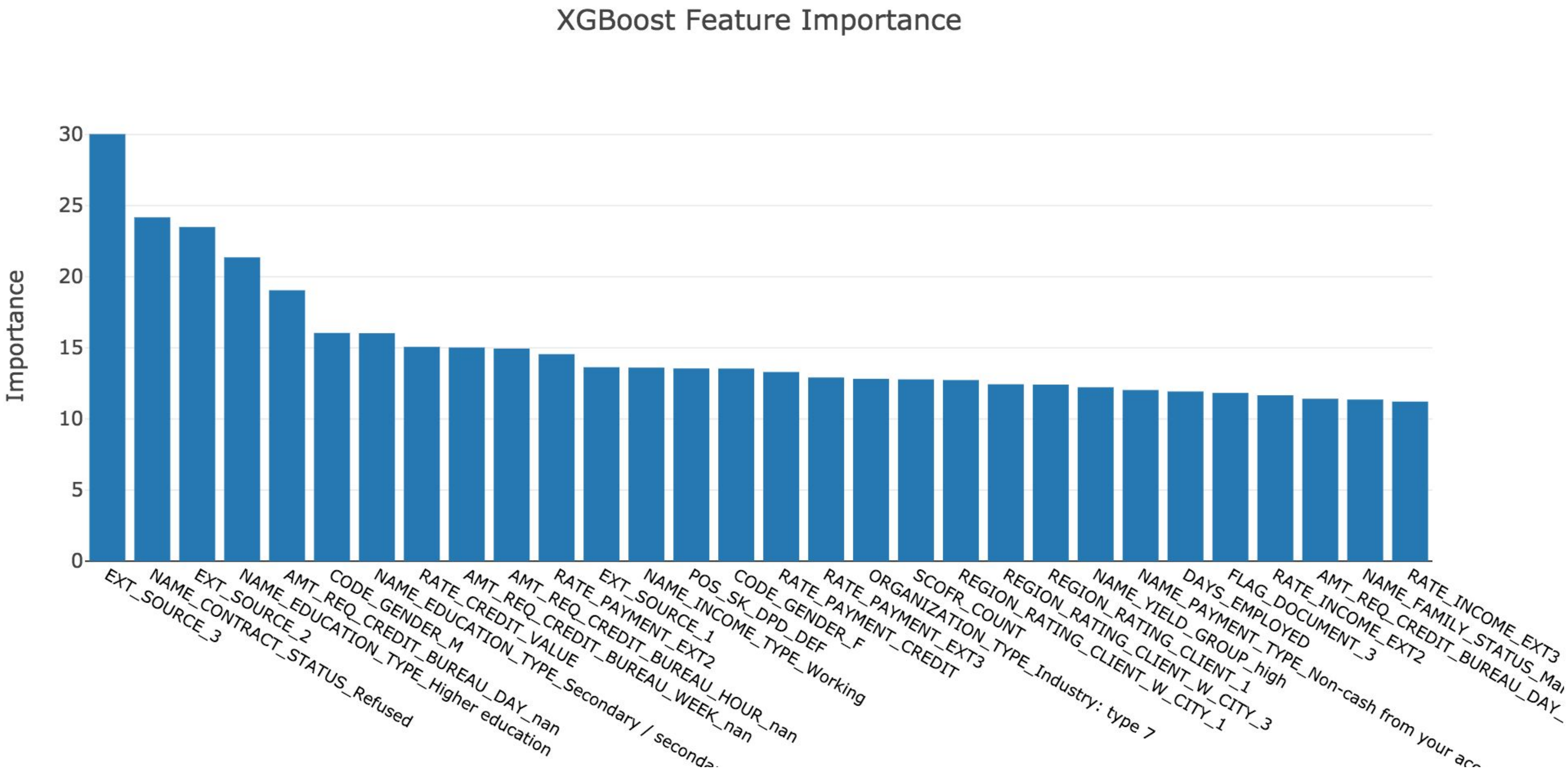
Problem

Baseline model

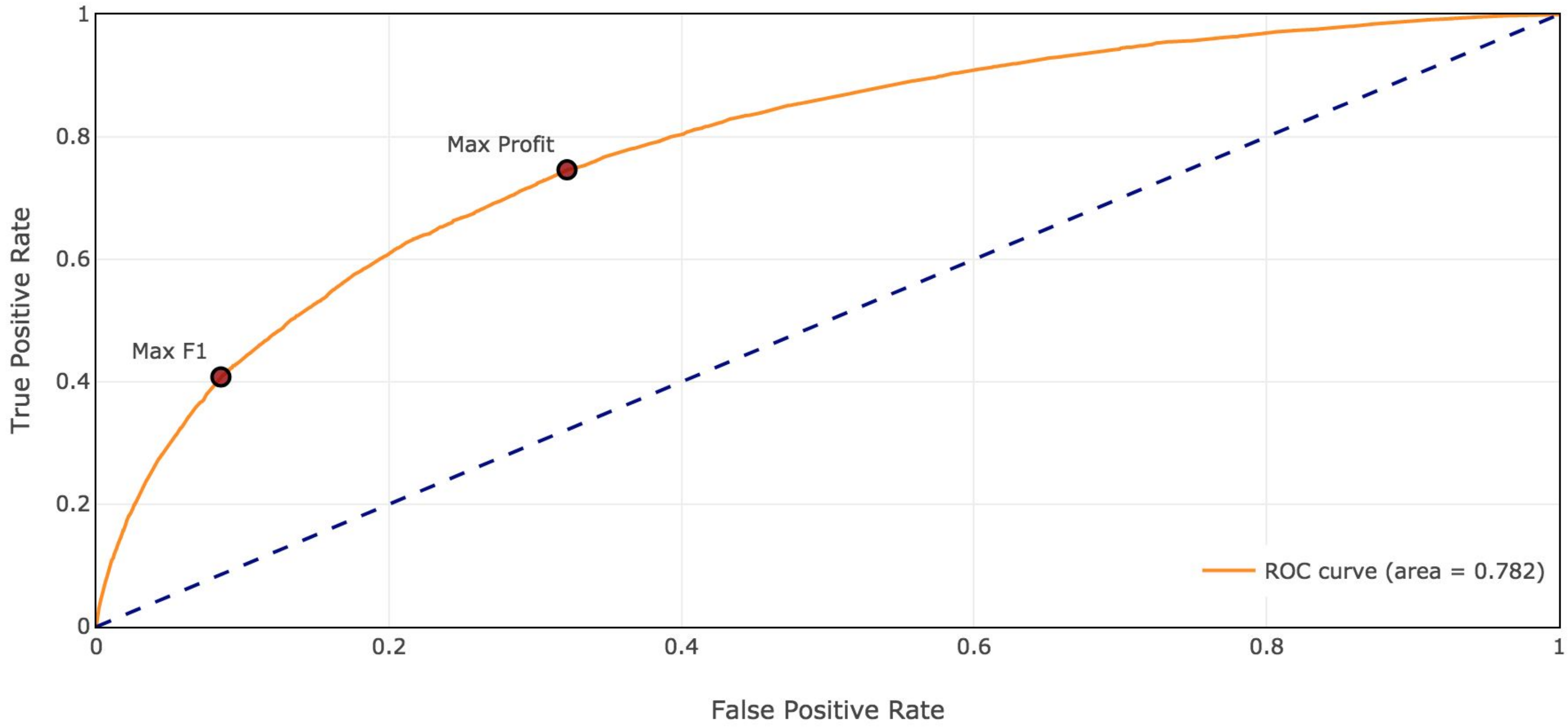
•Improved model

Future improvements

Complex model - Feature Importances



Complex Model - ROC Curve



Future Improvements

- Feature engineering. Use featuretools to try lots of features, ratios on a smaller dataset
- Use resampling to address class imbalance
- Stack / ensemble diverse algos
- Inspect false negatives for clues
- Read discussions and solutions to see how they achieved better AUCs

Choose Metrics You Care About ($)

- We used arbitrary $ values, loan-level profit/loss even better
- If metric you care about is continuous and differentiable, make a custom loss function
- We optimize MSE as a proxy for something like accuracy which we care about but is not a good objective - not differentiable, convex
- You can have the best R-squared in the world but you can’t spend it!

Conclusion

It works
Can be improved
Potential impact:
$Billions and $Billions

Questions?

Implementation

Tools

- Google Cloud Platform
- Postgres
- Pyscopg2 and sqlalchemy
- Plotly (and matplotib)
- XGBoost
- LightGBM
- CatBoost
- sklearn, pandas, numpy

Data source

- https://www.kaggle.com/c/home-credit-default-risk
- Home Credit: http://www.homecredit.net/