Kaggle Seasons #02

QUEEN MARY MACHINE
LEARNING SOCIETY

Last competition

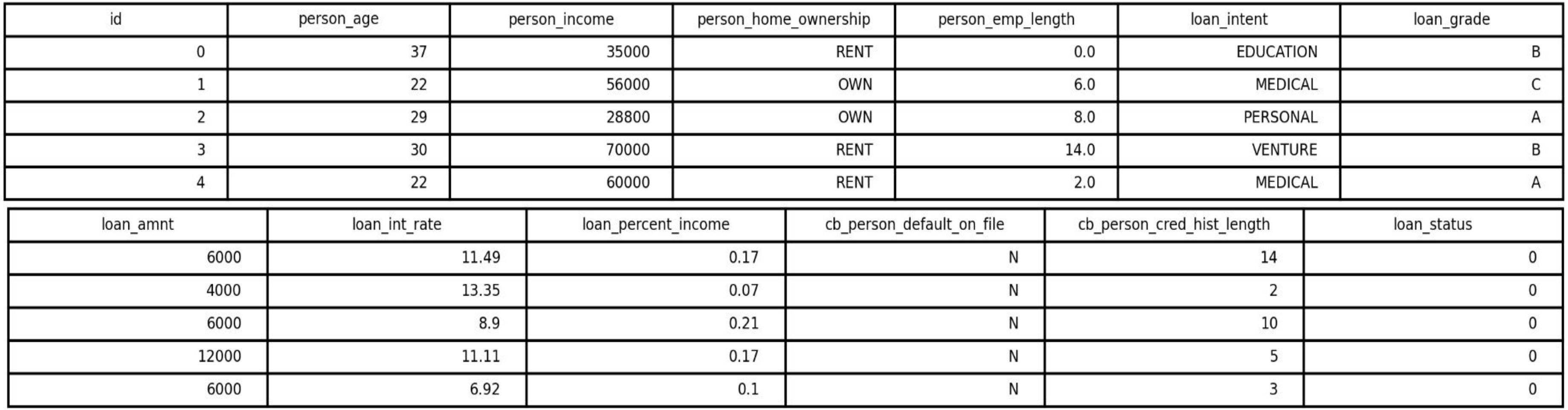
Loan Approval Prediction

Playground Series - Season 4, Episode 10

kaggle

Last competition

The data (first 5 rows)



id	person_age	person_income	person_home_ownership	person_emp_length	loan_intent	loan_grade	loan_amnt	loan_int_rate	loan_percent_income	cb_person_default_on_file	cb_person_cred_hist_length	loan_status
0	37	35000	RENT	0.0	EDUCATION	B	6000	11.49	0.17	N	14	0
1	22	56000	OWN	6.0	MEDICAL	C	4000	13.35	0.07	N	2	0
2	29	28800	OWN	8.0	PERSONAL	A	6000	8.9	0.21	N	10	0
3	30	70000	RENT	14.0	VENTURE	B	12000	11.11	0.17	N	5	0
4	22	60000	RENT	2.0	MEDICAL	A	6000	6.92	0.1	N	3	0

Last competition

The task

kaggle

Last competition

The task

...

kaggle

Last competition

The task

Predict loan status from the first 12 columns (for individuals not present in the dataset)

Last competition

General approaches

- *Data cleaning* (handle missing values and/or incorrect data)
- *Data enhancement* (find or construct similar datasets to increase data volume and/or help with data cleaning)
- *Encoding* (convert categorical variables into continuous variables so that they are parsable, or more well-interpretable, by our ML algorithm)
- *Normalisation* (standardise the scales of all of our variables so that arbitrary scale differences between variables don’t bias the learning process)
- *Model selection* (select the appropriate ML algorithm(s))
- *Hyperparameter tuning* (choose the best values for the parameters of our algorithm(s))

Last competition

General approaches (continued)

- *Feature engineering* (construct new variables, whether from the existing variables or from scratch, to feed into our ML algorithm(s))
- *Ensemble learning* (combine the knowledge gleaned by each of our ML algorithms)
- *Exploratory data analysis (EDA)* (understand the data intuitively with the help of statistics, tables, graphs, and other data visualisation techniques)
- *Evaluation* (determine how good our models are so we can track progress)
- *Subject-matter research* (gain information about the subject to contextualise our data)
- *Technical research* (research data science approaches relevant to our subject)

kaggle

Last competition

Specific approaches (examples)

- *Data cleaning*: Mean imputation, column dropping, row dropping, predictive modelling, duplicate removal, outlier removal
- *Data enhancement*: Data augmentation, synthetic data generation, oversampling
- *Encoding*: Label encoding, one-hot encoding, target encoding, binary encoding
- *Normalisation*: Z-score, L1, L2, min-max, robust (median-IQR) scaling
- *Model selection*: Logistic regression, Catboost, XGBoost, Random Forest
- *Hyperparameter tuning*: Manual search, grid search, random search, Optuna

kaggle

Last competition

General approaches (continued)

- *Feature engineering*: Principal component analysis (PCA), feature grouping
- *Ensemble learning*: Stacking, blending, hill climbing blending, bagging, voting ensemble
- *Exploratory data analysis (EDA)*: Summary statistics, box plots, histograms, correlation heatmap, missing value heatmap
- *Evaluation*: Train-test split, cross-validation, evaluation metrics: accuracy, recall, precision, AUC, (note: evaluation metric will be provided by Kaggle)
- *Subject-matter research*: Wikipedia, ArXiv, expert consultation
- *Technical research*: Kaggle competition discussion, Kaggle public notebooks

kaggle

Last competition

My approach

- *Data cleaning*: Mean imputation – replace missing values with the mean value of the corresponding variable
- *Data enhancement*: Append the original dataset, from which the competition dataset was synthesised, to the competition dataset
- *Encoding*: Label encoding – simple one-to-one mapping to integers
- *Normalisation*: Z-score normalisation: assume each variable is normally distribution and rescale the distribution to standard normal
- *Model selection*: Catboost, XGBoost, LGBM were my top-performing models
- *Hyperparameter tuning*: Manual search with reference to Kaggle public notebooks

kaggle

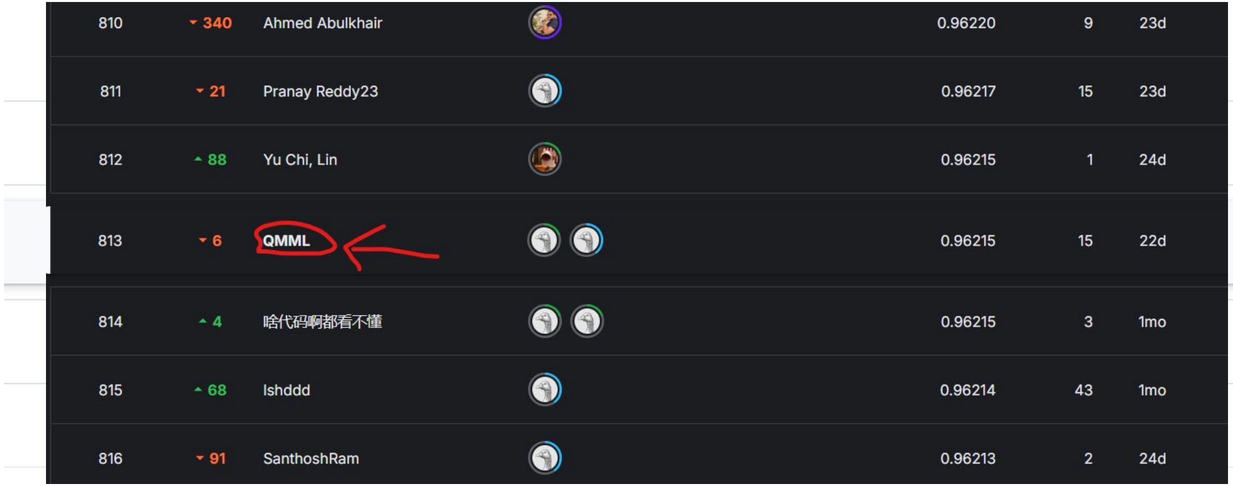
Last competition

My approach

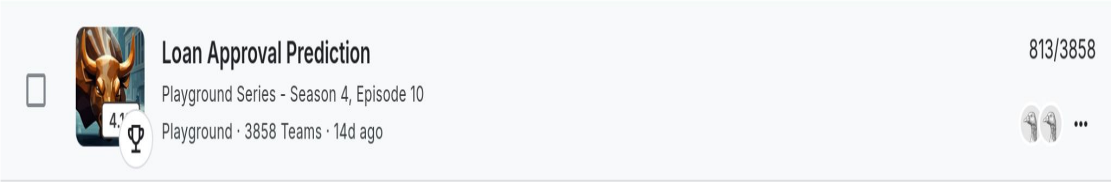
- *Feature engineering*: For CatBoost, use both the categorical and encoded versions of the categorical data
- *Ensemble learning*: Hill climbing blending: use simple linear regression with the hill climbing solver to combine predictions from my 3 models
- *Exploratory data analysis (EDA)*: Correlation heatmap, summary statistics
- *Evaluation*: Train-test split, stratified 5-fold cross-validation, evaluation metrics: AUC (required by the competition)
- *Subject-matter research*: Consultation with friend who works in finance
- *Technical research*: Kaggle competition discussion, Kaggle public notebooks

Last competition

Results



810	▾ 340	Ahmed Abulkhair	0.96220	9	23d
811	▾ 21	Pranay Reddy23	0.96217	15	23d
812	▴ 88	Yu Chi, Lin	0.96215	1	24d
813	▾ 6	QMML	0.96215	15	22d
814	▴ 4	啥代码啊都看不懂	0.96215	3	1mo
815	▴ 68	lshddd	0.96214	43	1mo
816	▾ 91	SanthoshRam	0.96213	2	24d



☐ Loan Approval Prediction

Playground Series - Season 4, Episode 10

Playground · 3858 Teams · 14d ago

813/3858

kaggle

Last competition

Results

Top 21%! Can you do better?

kaggle

Last competition

GOOD LUCK!

kaggle