CF Risk Score

https://github.com/moneykey/risk/tree/master/Saeede/CF%20Risk%20Score

Model Training Script: https://github.com/moneykey/risk/blob/master/Saeede/CF%20Risk%20Score/CF_Risk_leadgen_marketplace.ipynb

PMML - Leadgen: https://github.com/moneykey/risk/blob/master/Saeede/CF%20Risk%20Score/CF_risk_leadgen_mar2021.pmml

PMML - Marketplace: https://github.com/moneykey/risk/blob/master/Saeede/CF%20Risk%20Score/CF_risk_mar2021.pmml

Environment: https://github.com/moneykey/risk/blob/master/Saeede/CF%20Risk%20Score/environment.yaml

Introduction:

The CF risk scores were developed to come up with a full custom CF score to replace it with HD score, which was built based on, and for the most similar mk products to cf products in terms of APR . CF specific score should incorporate real experience to sharpen ROC curve.

Parameters:

- min insert_date: June 2019; max_insert_date: end of July 2020
- Target variable: Good and bad customers

Data Collection:

- CF Clarity data study is requested from Clarity for CF leads from June to Nov 2019 (from starting CF, to lunching HD risk Score) by providing them tracking numbers. ('CBWB_frm_Clarity_20200805.csv')
- The dataset is collected from clarity data study and features renamed to production names. (mapping_newClarityStack_dataStaudy_to_prod_Sep2020_version.ipynb)
- Duplicate columns removed and only the originated loans kept. (clarity_retro_databasenames.pkl)
- Raw product data corresponding to these loans collected from internal data bases. (reporting_cf.leads_accepted, datawork. mk_application , datawork.mk_whitepages , datawork.mk_clearrecenthistory , datawork.mk_clearinquiry , jaglms.lms_loan_header , jaglms.lms_base_loans , jaglms.lms_payment_schedules , jaglms.lms_customer_info_flat , jaglms.lms_payment_schedule_items , jaglms.lead_source)
- Getting application data + velocity metrics for the related leads.
- Only leadgen and marketplace leads in CF kept. (Channels: 'LeadGen-NM', 'LeadGen-M', 'LeadGen-LINC', 'Short Form', 'LendingTree', 'QuinStreet', 'CreditSesame')
- Normalizing true and false columns and converting them into numeric avoiding an added dimensionality when one hot encoding.
- Deleting columns with a high ratio of missing values that are not on the exception list.
- Duplicate columns removed.

Target variable labelling:

- The model is targeting the separation between good and bad loans
- A loan is considered bad if there was at least one miss payment within first 6 payments
- Otherwise a loan is considered good if there was not any miss payment within first 6 payments

Data Sources:

For the final models the following products have been chosen:

- CRH (Clear Recent History)
- CFI (Clear Fraud Insight)
- CI (Clear Inquiry)
- CCR (Clear Credit Risk)
- CAA (Clear Advanced Attributes)
- WP (White Pages)

Model development:

- Some features (bank and employee related) are set as null for marketplace leads as they are not available for them in hb1, when the scores are running (Such as: _delta_bankaccount_24h', '_delta_bankaccount_90d', 'bankaccount_to_bankaccount', 'bankaccount_to_dl', 'bankaccount_to_email', 'bankaccount_to_homephone', 'bankaccount_to_income', 'bankaccount_to_otherphone', 'bankaccount_to_ssn', 'bankyears', 'dl_to_bankaccount', 'email_to_bankaccount', 'emp_start','empstate', emptype', 'empyears', 'homephone_to_bankaccount', 'ipaddress', 'n_months_bank', 'n_months_emp', 'routingnumber', 'ssn_to_bankaccount', 'MKA_IsDiffState_EmpState', 'MKA_DaysSinceEmptart', 'MKA_AgeEmpstart', 'MKA_AgeEmpstartInRange', 'MKA_RatioBA24h90d', 'first_seen_bankaccount', 'ip_state_count_2', 'ip_state_count_5')
- Features like age, cfi score, seasonality, payfrequency are removed. (Discriminative or double-using features)
- Last month of data (July 2020), and last 1.5 month before lunching hd (2019-10-01 to 2019-11-15) score are kept as out of sample testing.
- After separating out of sample set, data is split using a random simple split of 0.8 Training; 0.1 Validation and 0.1 Testing
- Due to different nature of marketplace and leadgen leads and to come up with a better performance, one model is built for leadgen which is just trained by leadgen leads (**'mk_CF_risk_leadgen_score'**), while another model is built for marketplace, trained by leadgen and marketplace leads as we didn't have enough marketplace leads (**'mk_CF_risk_marketplace_score'**)
- The 1st iteration of xgboost model is trained with the full dataset (all features) and **Bayesian Hyperparameter optimization**. Then a model with best set of parameters are trained to detect features with zero importance and drop them from the feature list.

- Performance of the model is measured on test set and compared with the performance of **HD Risk Score** out of sample test set(July 2020) as the base line. Based on **FPD**, AUC improved from **0.53** to **0.59** for **leadgen** and from **0.57** to **0.63** for **marketplace**, for same lead.
- If performance is approved and model is deemed ready for production use. The out of sample testing+ train + vald + test datasets are merged and the final models are trained.

Deployment in production:

- Upon validation and testing the model is deployed in production
- hotbox scores (var_name **'mk_CF_risk_leadgen_score'** and **'mk_CF_risk_marketplace_score'** from **jaglms.mk_lead_data30**) are compared with java test score and in case of any difference, features are compared to debug.
- **'mk_CF_risk_marketplace_score'** deployed on stream CFI Ext Shortform RO(49) and **'mk_CF_risk_leadgen_score'** on CFI Leads (46)

Hints for next versions:

- Following features are not available for marketplace in hb1: dow (day of week for next payment), /xml-response/inquiry/date-of-next-payday, /xml-response/inquiry/bank-account-zeros