Lenus Case Study

– Pallavi Ravikiran Bhat

Objective: To analyze the dataset and determine the most important features to predict conversion

Approach: Utilize heuristic methods, visualizations, statistical-tests, and feature selection methods (filter and embedded selections) assess and conclude importance of each feature

Initial Exploration and Pre-processing

Data Overview

- Raw data: 891 rows x 10 columns
- Primary key: Data is unique at customer ID granularity with no duplicity
- Only 204 of the customers have a credit account ID and 46 of these account IDs are one-to-many mapped to customer ID, this key is discarded from analysis
- Credit account ID has only 2'Unnamed: 0' and 'id' dropped since they are not required for analysis

Types of Features

- **Continuous:** age, initial fee level
- **Nominal Categorical:** customer segment, gender, branch
- **Ordinal Categorical:** related customers, family size (these are treated as continuous for analysis)
- **Dependent binary variable:** converted

Outliers / Noise

- Age values start from 0, in the absence of business context, age 1 and above are considered valid and the values between 0 and 1 are rounded up to 1.

Missing Values

- Age has 19.9% missing values (177 rows). Distributions of all features were compared for Age=null and Age=not null, there was no pattern found to the missingness, so it is assumed that the feature is Missing is at Random (MAR).
- These rows are not dropped although they are above the usual threshold for imputation (5-10% of the total records) since the volume of the data is already a sample. Imputation methods tried:
  - **Mean Imputation**: This maintains the sample mean, but it reduces variance of the distribution and introduces some bias
  - **Ffil Imputation**: This uses the last valid observation to fill the null values. Although this seemed to reduce concentration of datapoints to one value, it still causes distortion

- **Median Imputation**: Performs best in terms of minimizing the variance distortion of the right skewed age, standard deviation drops from 14.53 to 13.02 after imputation

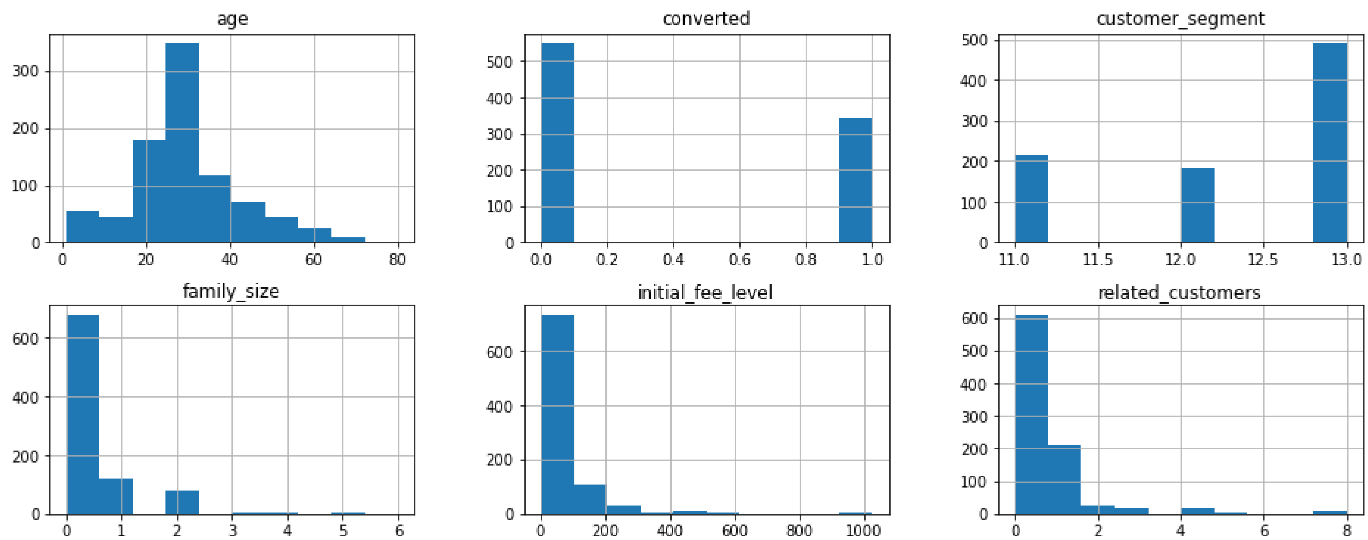
- Branch has only 2 rows with missing values which are imputed using the mode 'Helsinki'

Assumptions

1. Ages 1 or above are considered valid, only age <1 treated as noise
2. Imputation of the 20% missing records of Age does not cause significant distortion

Exploratory Data Analysis

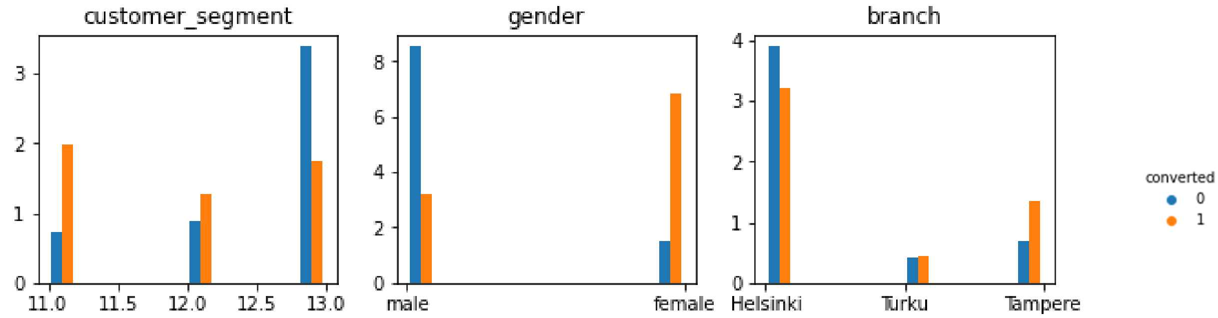
Univariate Analysis – Numerical



- Age is roughly right skewed with normal distribution
- Dependent variable 'converted' does not have significant class imbalance, can be resolved by stratification in train-test split
- Some features (related customers, family size, initial fee level) have exponential decay distributions, these are not treated through log transformation since:
    1. They are not continuous, but discrete
    2. Log transformation is applied only when the feature is to be normalized. However, in this analysis, only tree-based algorithms are considered for embedded methods of feature selection which do not require normalized features.
- Age has fairly symmetrical distribution and only slightly thicker tail than normal distribution, Skewness and Kurtosis shown below fall under acceptable ranges for normality assumption

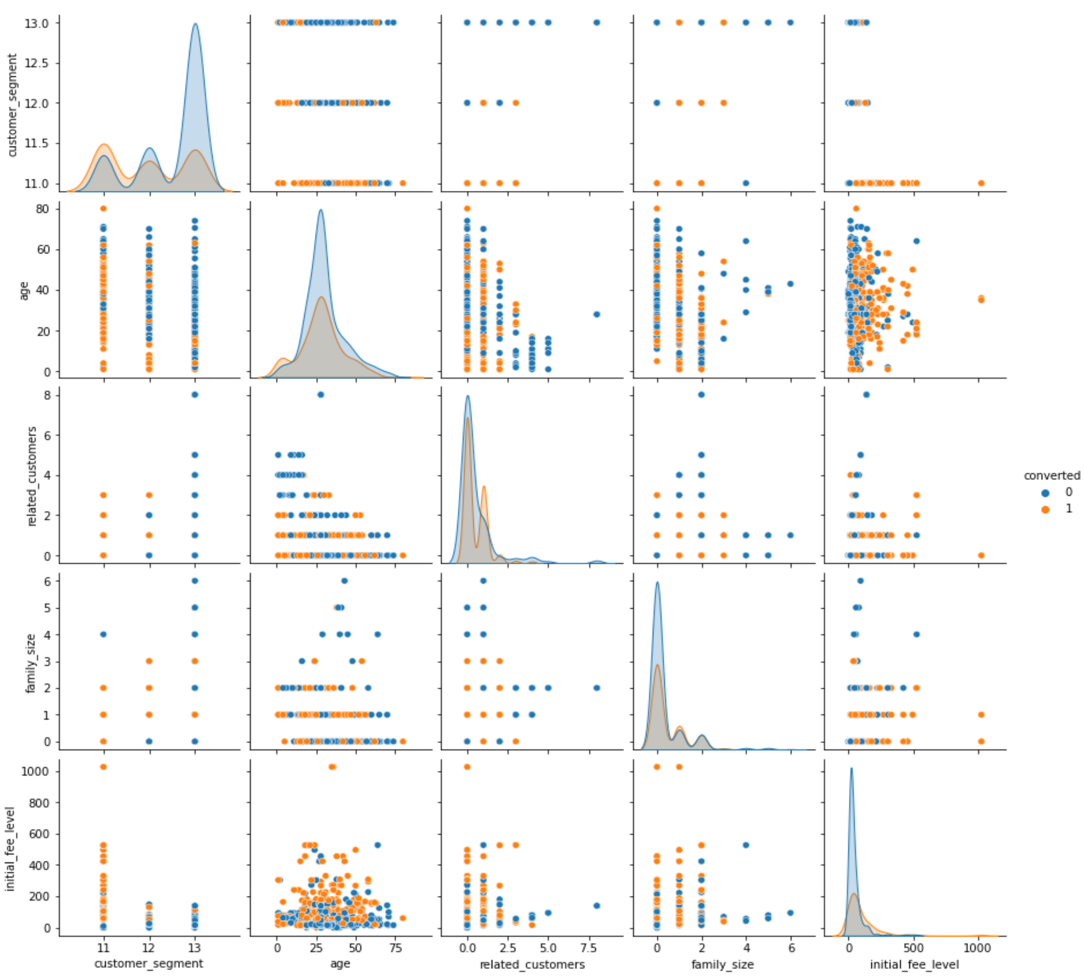
```
skew        0.389108
kurtosis    0.178274
```

Univariate Analysis – Categorical



- Segment 11 leads have very high propensity for conversion while segment 13 leads have low conversion rate, segment 12 might not be useful in discriminating between the classes of y
- product / service seems to be targeted at females who are converting 3 out of 4 times
- Helsinki: most rejects, Turku: worst conversion rate, Tampere: best conversion rate

Bivariate Analysis – Numerical



- Leads are most likely to convert if they have 1 related customer
- Under 25 leads with 3 or more related customers are most likely to not convert

- Family size and related customers seem somewhat correlated, with dense non-conversions for higher values (tails), relationship to be further tested statistically
- Leads with initial fee level higher than the inter-quartile range are most likely to not convert, initial fee level of ~1000 could be potential outlier

Pearson's Correlation

	age	related_customers	family_size	initial_fee_level
age	1.000000	-0.233328	-0.172394	0.096688
related_customers	-0.233328	1.000000	0.414838	0.159651
family_size	-0.172394	0.414838	1.000000	0.216225
initial_fee_level	0.096688	0.159651	0.216225	1.000000

- X-X correlation: Moderate correlation between family size and related customers, remaining features show low correlation with each other, so they can be retained for training
- X-y correlation: The correlation between X features and target is low to medium. There is no strong correlation with one particular feature that could have been isolated for evaluation, hence all features are retained
- Correlation is not the most suitable technique for a binary target, hence other measure explored

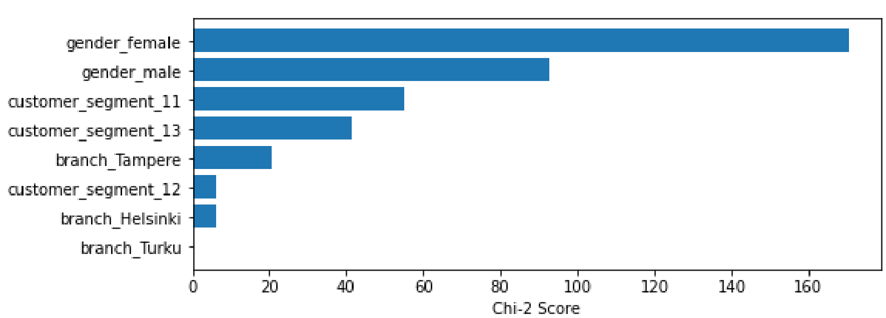
Variance Inflation Factor (VIF)

Variable	VIF
age	1.092125
related_customers	1.262927
family_size	1.258119
initial_fee_level	1.082116

- Correlated continuous features observed from the correlation matrix are cross-checked using Variance Inflation Factor (VIF), a threshold of 5 is set as an indicator of multicollinearity
- All features fall under the threshold, hence multicollinearity is absent

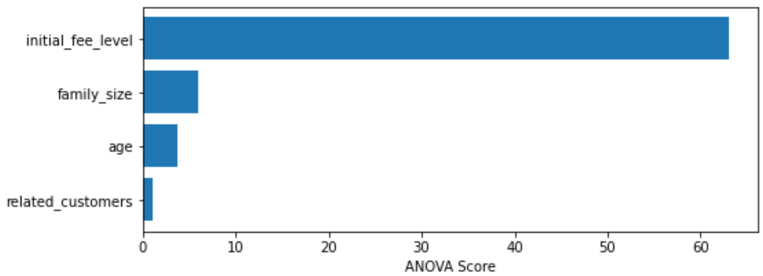
Filter Methods for Feature Selection

Chi-Square Test - Categorical Features



- This test is used to quantify feature importance of categorical features for classification models.
- A small score implies independence with respect to target, while a large score implies non-random relation to the target, and likely important.
- One class of each one-hot-encoded feature is to be dropped: Gender male, customer segment 12 and Branch Turku are removed from analysis due to lowest Chi-square scores among the other classes of the same feature

ANOVA Test - Continuous Features



- ANOVA is used to assess how well continuous features discriminate between the two classes of the binary dependent variable
- 'Related customers' is dropped from analysis since it has a very low F-test score and this information is somewhat contained in 'family size' due to their Pearson's correlation (41%)

Feature Selection Summary

(Using Visualizations, Heuristic and Statistical Methods detailed above)

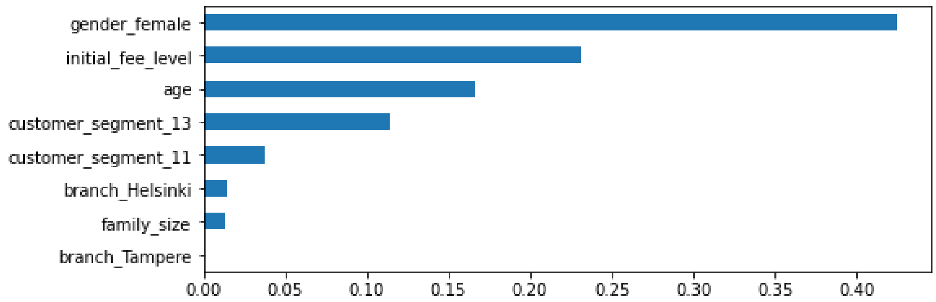
Feature	Type	Status	Reason
customer_id	ID	dropped	irrelevant
credit_account_id	ID	dropped	irrelevant
age	Continuous	retained	decent ANOVA score
family_size	Ordinal (discrete)	retained	decent ANOVA score
related_customers	Ordinal (discrete)	dropped	very low ANOVA score, and correlated to family_size
initial_fee_level	Continuous	retained	highest ANOVA, good predictor
customer_segment_11	Categorical	retained	good chi-2 score
customer_segment_12	Categorical	dropped	very low chi-2 score
customer_segment_13	Categorical	retained	good chi-2 score
gender_male	Categorical	dropped	lower chi-2 score than class female, redundant class
gender_female	Categorical	retained	highest chi-2 score
branch_Helsinki	Categorical	retained	decent chi-2 score
branch_Tampere	Categorical	retained	good chi-2 score
branch_Turku	Categorical	dropped	lower chi-2 score than the other two branches, redundant class

Note: Prediction power can be quantified and analyzed among the retained features using embedded feature selection through tree-based algorithms as explained in the next section.

Embedded Methods for Feature Selection

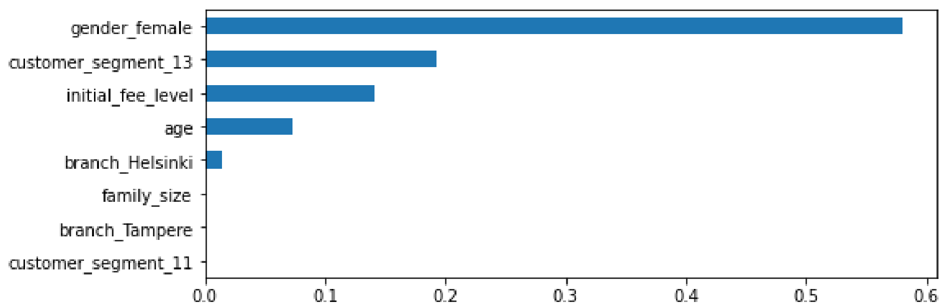
Iteration 1: Decision Tree

Parameters: `criterion=gini, max_depth=10, min_samples_split=2, min_samples_leaf=2`



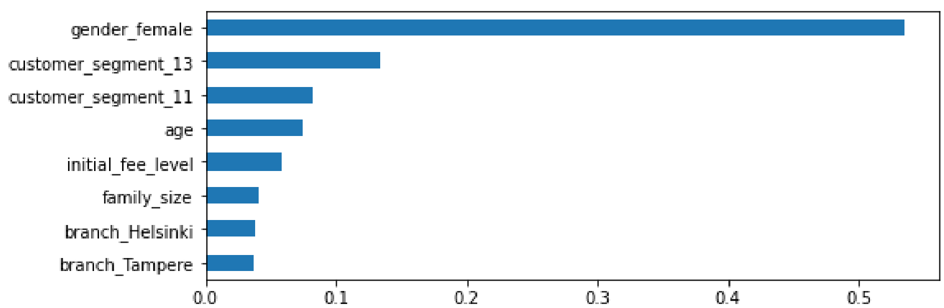
Iteration 2: Best Decision Tree model from GridSearchCV

Parameters: `criterion=entropy, max_depth=3, min_samples_leaf=2, min_samples_split=5`



Iteration 3: Best XGBoost model from GridSearchCV

Parameters: `gamma=3, learning_rate=0.2, max_depth=4, all other parameters=default`



Conclusions

- Selected subset of features obtained through reduction using statistical methods are fed into tree-based classification algorithms for a second layer of selection
- Embedded methods use Gini Index / Entropy for splitting criteria in tree-based algorithms, these models are iteratively tuned to analyze and compare feature importance of each predictor
- Feature importance captures the decrease in node weighted by the probability in reaching that node.

- Best to worst ranked features based on the magnitude of their importance from the above iterations and statistical tests:

Feature	Importance	Cumulative Importance
gender_female	55%	55%
customer_segment 13	17%	72%
initial_fee_level	10%	82%
age	7%	89%
customer_segment 11	5%	94%
branch_Helsinki	3%	97%
family_size	2%	99%
branch_Tampere	1%	100%
related_customers	0%	-
customer_segment_12	0%	-
gender_male	0%	-
branch_Turku	0%	-

Note: Python notebook attached with the report for reference