Clustering

2024-11-24

Table of contents

Clustering

Normalized information about building where the client lives

Missing data

We conducted Principal Component Analysis (PCA) on the building-related features after careful preprocessing. The initial analysis of missing values revealed significant data gaps, with missing rates ranging from 0% to 69.87%. Common areas (COMMONAREA_AVG/MODE/MEDI) showed the highest missing rate at approximately 69.87%, while administrative features like EMERGENCYSTATE_MODE and WALLSMATERIAL_MODE had no missing values.

	Variable	Missing_Count	Missing_Percent
1	COMMONAREA_AVG	214865	69.87230
2	COMMONAREA_MODE	214865	69.87230
3	COMMONAREA_MEDI	214865	69.87230
4	NONLIVINGAPARTMENTS_AVG	213514	69.43296
5	NONLIVINGAPARTMENTS_MODE	213514	69.43296
6	NONLIVINGAPARTMENTS_MEDI	213514	69.43296

Given the high proportion of missing values, we implemented a **50% missing rate threshold** for variable selection. This preprocessing step reduced our feature set from 47 original variables to 7 key variables, ensuring more reliable analysis while maintaining data quality.

Principal Component Analysis

The PCA results were notably effective, with the first two principal components explaining 91.97% of the total variance: - Principal Component 1 (PC1) accounts for 51.47% of the variance - Principal Component 2 (PC2) accounts for 40.50% of the variance

The first principal component (PC1) is primarily characterized by building structure features, with the highest loadings from:

1. FLOORSMAX_MEDI (0.484)

2. FLOORSMAX_AVG (0.484)

3. FLOORSMAX_MODE (0.482)

4. TOTALAREA_MODE (0.382)

5. YEARS_BEGINEXPLUATATION_AVG (0.229)

The second principal component (PC2) is predominantly influenced by building age and similar characteristics:

1. YEARS_BEGINEXPLUATATION_AVG (0.532)

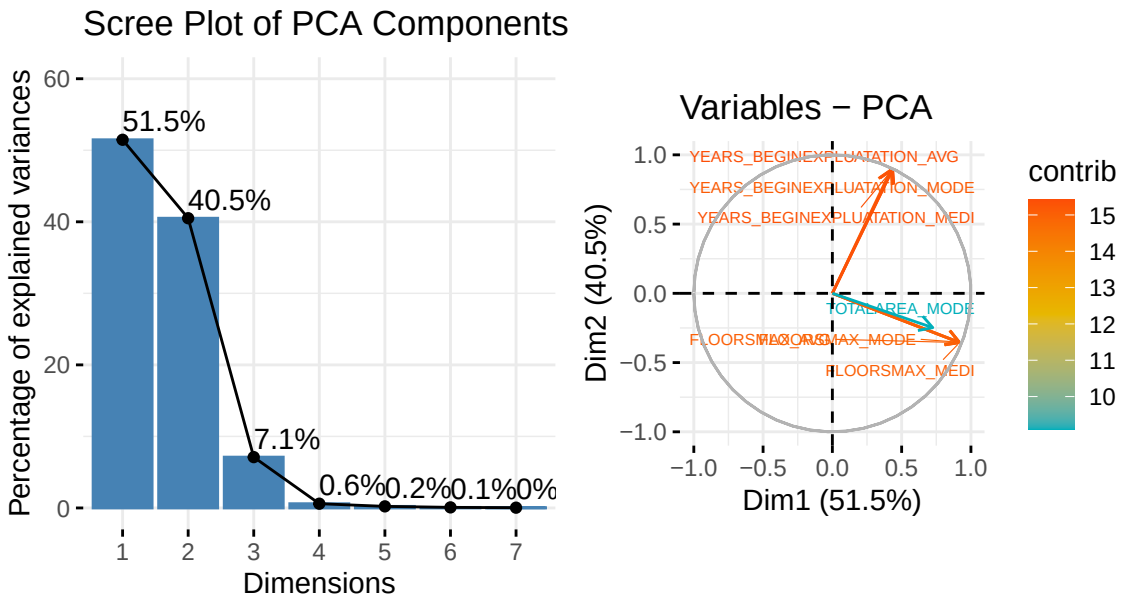
2. YEARS_BEGINEXPLUATATION_MEDI (0.531)

3. YEARS_BEGINEXPLUATATION_MODE (0.529)

4. FLOORSMAX_MEDI (0.211)

5. FLOORSMAX_AVG (0.211)

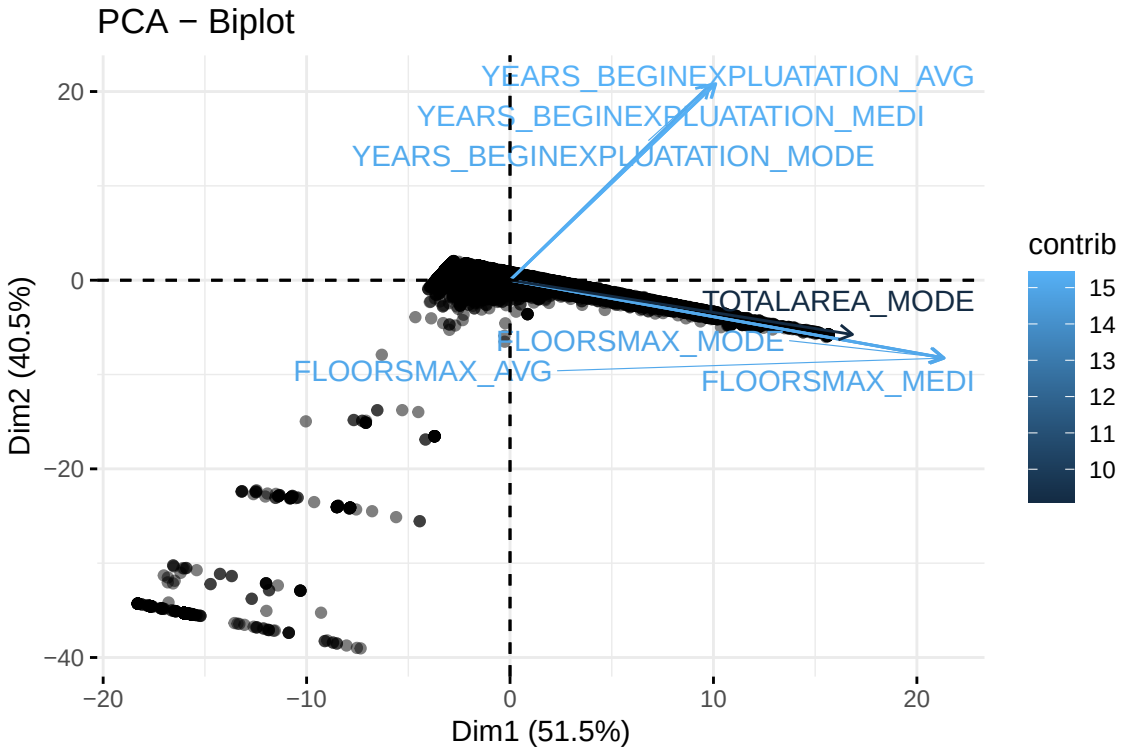
```
Cumulative proportion of variance explained:
[1] 0.5146554 0.9196713 0.9907191 0.9966518 0.9988456 0.9996315 1.0000000
```




```
Suggested number of components (80% variance explained): 2
```

Given the high cumulative explained variance (91.97%) with just two components, this dimensional reduction effectively captures the essential patterns in the building-related features while significantly simplifying the feature space. PC1 appears to represent the physical scale of buildings (dominated by floor and area measurements), while PC2 primarily captures the temporal aspects (building age and exploitation period)."

This analysis suggests that the complex building-related information can be effectively summarized using just these two principal components while retaining over 90% of the original information.



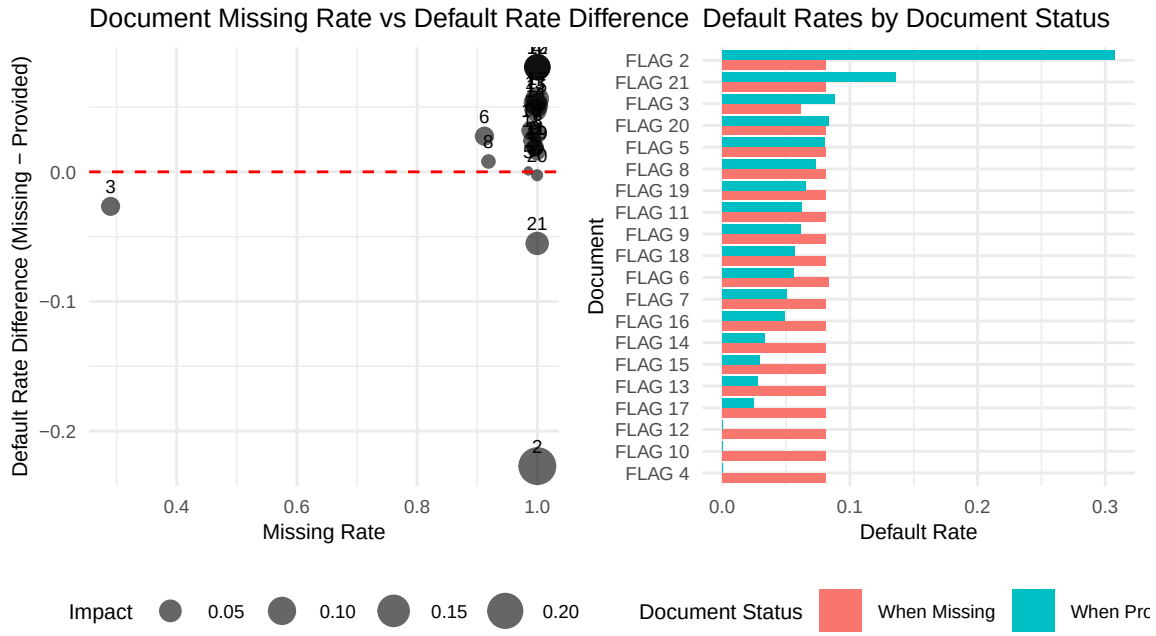
Provided Documents

Upon examining the relationship between document provision status and default rates, we consistently observe that when documents (most of them) are provided, default rates are significantly lower than when they are missing. This negative association is particularly pronounced in documents like FLAG 4, where the difference in default rates (missing vs. provided) is the most substantial. The systematic nature of this relationship and its considerable effect size indicate that these document flags have significant predictive power for default risk.

Therefore, we recommend retaining these original binary variables in their current form for the modeling process, rather than applying any feature aggregation or transformation. This choice is further supported by the clear interpretability of these features, which is essential for model transparency and practical business applications.

	not_provided_count	not_provided_rate	default_rate_when_missing
FLAG_DOCUMENT_4	307486	0.9999187	0.08073538
FLAG_DOCUMENT_10	307504	0.9999772	0.08073066
FLAG_DOCUMENT_12	307509	0.9999935	0.08072934
FLAG_DOCUMENT_17	307429	0.9997333	0.08074385
FLAG_DOCUMENT_13	306427	0.9964749	0.08091650
FLAG_DOCUMENT_15	307139	0.9987903	0.08079078

	default_rate_when_provided	default_rate_diff
FLAG_DOCUMENT_4	0.00000000	0.08073538
FLAG_DOCUMENT_10	0.00000000	0.08073066
FLAG_DOCUMENT_12	0.00000000	0.08072934
FLAG_DOCUMENT_17	0.02439024	0.05635360
FLAG_DOCUMENT_13	0.02767528	0.05324122
FLAG_DOCUMENT_15	0.02956989	0.05122089

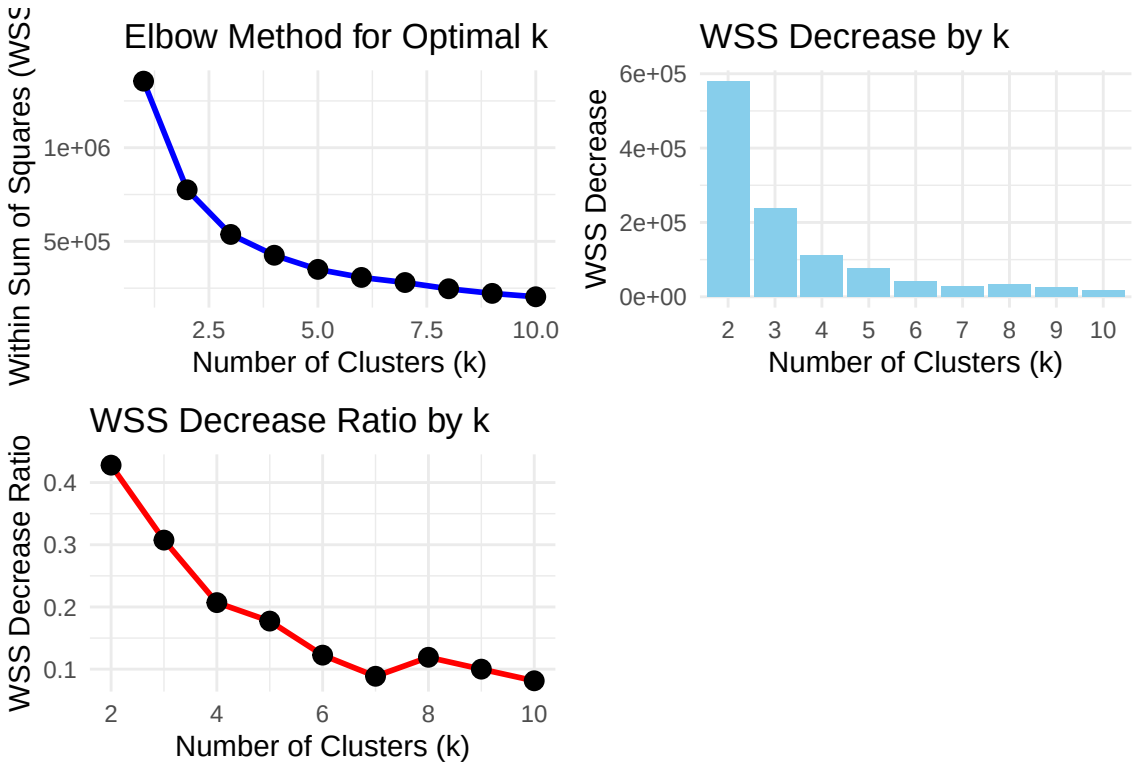


Previous Application

Historical application patterns from previous application data are strong predictors of future default behavior. To effectively incorporate this information, we aggregate multiple historical applications ('SK_ID_PREV') linked to each current application ('SK_ID_CURR') using **a time-weighted approach**. This method emphasizes recent behavior while preserving the relevance of historical patterns through exponential time decay weighting, capturing the temporal evolution of the client's credit behavior and risk profile.

Based on the order of `DAYS_DECISION` (from recent to old), we generate time weights using the formula (time_weight $=$ $\exp\left(-\frac{|\text{DAYS\_DECISION}|}{365}\right)$). Using these weights, we aggregate data by `SK_ID_CURR` to create four variables: `weighted_approval_rate`, `weighted_credit_ratio`, `recent_approval`, and `trend_credit_ratio`. This approach allows us to quantify the influence of historical application behavior on current credit risk assessments.

K-means clustering was applied to segment application behaviors. The optimal number of clusters was determined using **the elbow method**, which examined both absolute WSS and its decrease ratio. The analysis indicated that **k=5** offers the most efficient segmentation, with minimal marginal improvements beyond this point, suggesting five distinct behavioral patterns in the historical application data.



The K-means clustering analysis revealed five distinct patterns in historical application behavior:

Cluster 1 (33,044 customers):

- Approval Rate: 0.865 (High)
- Credit Ratio: 0.994 (Balanced)
- Recent Approval History: 0.873 (Strong)
- Represents stable, consistent borrowers.

Cluster 2 (47,908 customers - largest group):

- Approval Rate: 0.974 (Very High)
- Credit Ratio: 1.13 (Above Average)
- Recent Approval: 0.996 (Excellent)
- Represents prime customers with strong credit profiles.

Cluster 3 (11,501 customers):

- Approval Rate: 0.978 (Highest)
- Credit Ratio: 0.743 (Conservative)
- Recent Approval: 0.987 (Very Good)
- Represents cautious borrowers who consistently get approved but apply for smaller loans.

Cluster 4 (127,765 customers):

- Approval Rate: 0.288 (Very Low)
- Credit Ratio: 1.01 (Average)
- Recent Approval: 0.000376 (Poor)
- Represents high-risk applicants with consistent rejections.

Cluster 5 (118,639 customers):

- Approval Rate: 0.848 (Good)
- Credit Ratio: 0.996 (High)
- Recent Approval: 1.0 (Perfect)
- Represents recently improved borrowers with an excellent recent history.

The clustering effectively separates customers based on their historical performance, showcasing clear distinctions in approval rates and credit utilization patterns across segments.

```
[1] "Cluster Summary:"
```

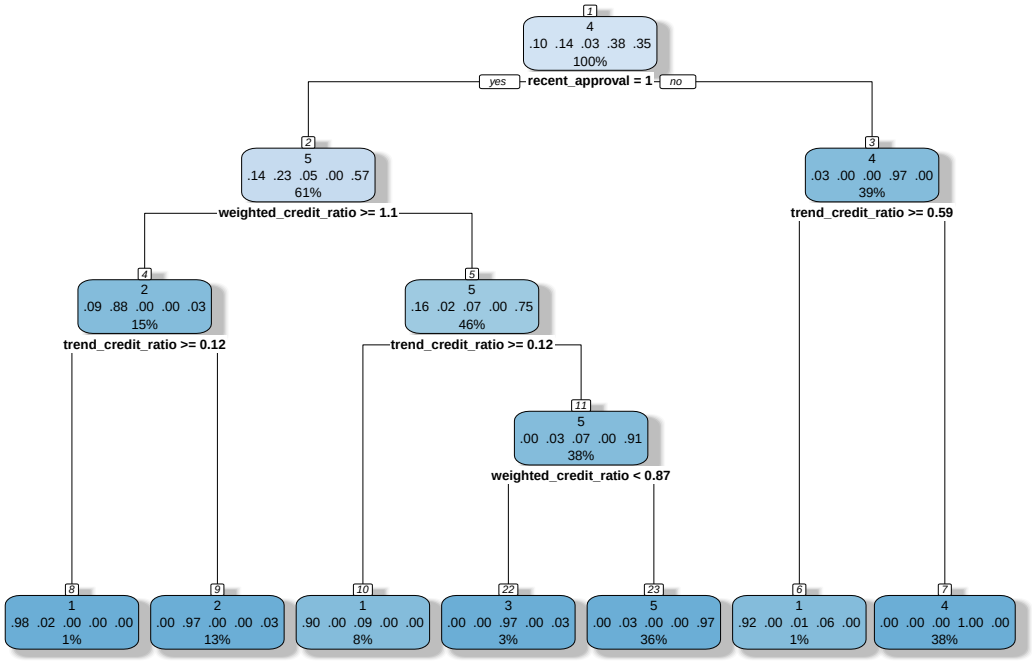
```
# A tibble: 5 x 6
  cluster   size avg_weighted_approval avg_weighted_credit avg_recent_approval
    <int>  <int>                 <dbl>               <dbl>               <dbl>
1       1  33044                 0.865               0.994           0.873
2       2  47908                 0.974               1.13            0.996
3       3  11501                 0.978               0.743           0.987
4       4 127765                 0.288               1.01            0.000376
5       5 118639                 0.848               0.996           1
# i 1 more variable: avg_trend <dbl>
```

Using historical application data, we identified five distinct customer segments through K-means clustering. The decision tree analysis indicates that recent approval status is the primary differentiator, followed closely by credit ratio trends. The segments are as follows:

- **High-Risk Segment (Cluster 4)**: 37.7%
- **Recently Improved (Cluster 5)**: 35.0%
- **Prime Customers (Cluster 2)**: 14.1%

- **Stable Performers (Cluster 1)**: 9.75%
- **Conservative Borrowers (Cluster 3)**: 3.39%

This segmentation provides a clear framework for understanding customer risk profiles based on their historical application patterns.




```
# A tibble: 5 x 9
  cluster  count pct_total avg_approval_rate avg_credit_ratio
    <int>  <int>     <dbl>             <dbl>            <dbl>
1       4 127765     37.7              0.288            1.01
2       5 118639     35.0              0.848            0.996
3       2  47908     14.1              0.974            1.13
4       1  33044      9.75             0.865            0.994
5       3  11501      3.39             0.978            0.743
# i 4 more variables: recent_approval_rate <dbl>, avg_trend <dbl>,
#   sd_approval_rate <dbl>, sd_credit_ratio <dbl>
```