# Amount of Children Prediction

Using A Credit Card approval Dataset for Machine Learning

|  |  |
| --- | --- |
| Team Members: | ID: |
| Dvir Iluz | 311447668 |
| Shulamit Elgrabli | 208565515 |
| Yitzhak Vider | 315963488 |

**Project Goal:**  
Using machine learning to predict the number of children a bank customer.

**Project Motivation:**  
Helping the bank customize their marketing and promotions for credit cards that fit the needs of specific groups of customers. It can also help the bank predict which customers may have trouble paying their credit card bills, so the bank can set credit limits appropriately. In addition, we will try to look for groups with similar characteristics, to match them with relevant business proposals.

**Methodology:**

**The data was taken form Kaggle website:** [**https://www.kaggle.com/datasets/rikdifos/credit-card-approval-prediction**](https://www.kaggle.com/datasets/rikdifos/credit-card-approval-prediction) **and has 438557 rows (examples).**

Data: we will use data about bank client, with the following parameters:

|  |  |  |  |
| --- | --- | --- | --- |
| **Feature name** | **Explanation** | **Type/ Range** |  |
| ID | Client number | numeric |  |
| CODE\_GENDER | Gender | Category | M/F |
| FLAG\_OWN\_CAR | Is there a car | Category | Y/N |
| FLAG\_OWN\_REALTY | Is there a property | Category | Y/N |
| AMT\_INCOME\_TOTAL | Annual income | numeric |  |
| NAME\_INCOME\_TYPE | Income category | Category |  |
| NAME\_EDUCATION\_TYPE | Education level | Category |  |
| NAME\_FAMILY\_STATUS | Marital status | Category |  |
| NAME\_HOUSING\_TYPE | Way of living | Category |  |
| DAYS\_BIRTH | Birthday | numeric |  |
| DAYS\_EMPLOYED | Start date of employment | numeric |  |
| FLAG\_MOBIL | Is there a mobile phone | binary | 1/0 |
| FLAG\_WORK\_PHONE | Is there a work phone | binary | 1/0 |
| FLAG\_PHONE | Is there a phone | binary | 1/0 |
| FLAG\_EMAIL | Is there an email | binary | 1/0 |
| OCCUPATION\_TYPE | Occupation | Category |  |
| **CNT\_FAM\_MEMBERS** | **Family size** |  | **numeric (1-20)** |

**Method**

The project is split into two parts. The first part focuses on classification models to determine the number of children in a family. The second part uses the K-MEAN model to classify people into different groups and understand their unique characteristics. This analysis helps the bank make informed decisions.

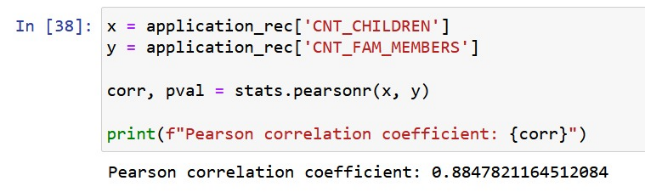
We used the following Machine Learning classification Algorithms for the predation:

* KNN
* Logistic Regression
* Random Forest
* Gradient Boosting
* K-Means (Clustering)

**Data Preprocessing:**

First, we loaded the data, and we prepared the data for the model:

1. We calculated the correlation between 'CNT\_CHILDREN' and 'CNT\_FAM\_MEMBERS' using the Pearson correlation coefficient.

****

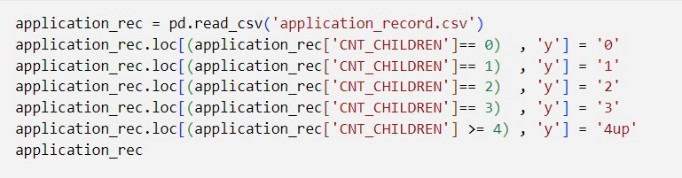
2. Converted the categorical features 'CODE\_GENDER', 'FLAG\_OWN\_CAR' and 'FLAG\_OWN\_REALTY' into binary values.

3. Created the 'employment\_status' variable using the 'DAYS\_EMPLOYED' feature.

4.1 Transformed the 'CNT\_CHILDERN' feature into a categorical variable based on the number of children and saved the values into the variable ‘Y’.



4.2 We created another python file, divided the data equally according to the number of children and ran the models.



5. Dropped unnecessary features like 'CNT \_CHILDERN', 'DAYS\_BIRTH', 'DAYS\_EMPLOYED', 'CNT\_FAM\_MEMBERS', and 'ID' from the dataset.

**Classification**

**In the next step we trained the model:**

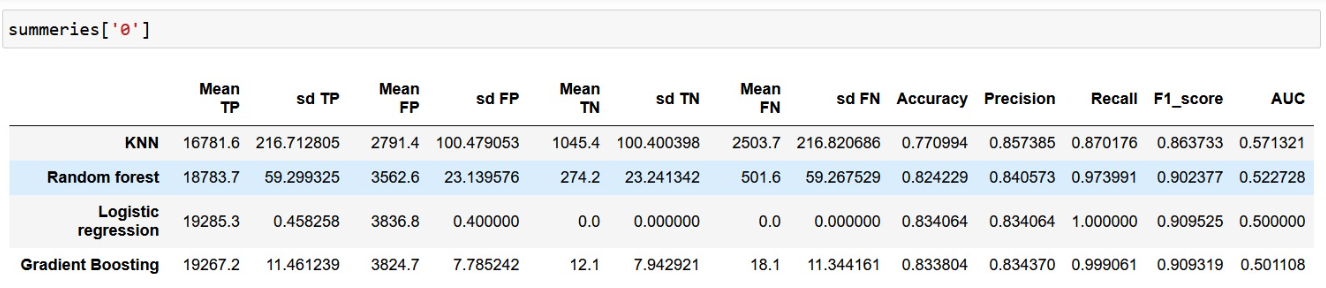
1. Used the Stratified K-Fold cross-validation technique with 10 folds to evaluate model performance.

2. Used one-hot encoder on categorical features and standardized numerical features**.**

3. We used the following Machine Learning classification Algorithms for the predation:

* KNN
* Logistic Regression
* Random Forest
* Gradient Boosting

4. Calculated accuracy, recall, precision, and F1-score for each model using the test dataset and saved them separately. We calculated this to each group separately. For example,



**Classification Conclusions:**

Table 1 demonstrated better performances compared to Tables 2, 3, 4, and 5, with higher precision, recall, and F1-score values. The models in Table 1 had moderate accuracy and showed relatively stronger capabilities in identifying positive instances. However, all tables displayed room for improvement, as the models struggled with low recall and inconsistent AUC values. Further optimization is needed to enhance the models' accuracy and overall predictive power. To improve the model's performance, we aimed to balance the data division based on the variable Y, representing the number of children in the family. Unfortunately, no improvement was found in the indices, and it even seems that there was a deterioration in performance.

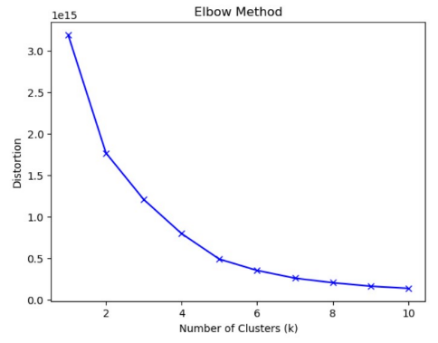
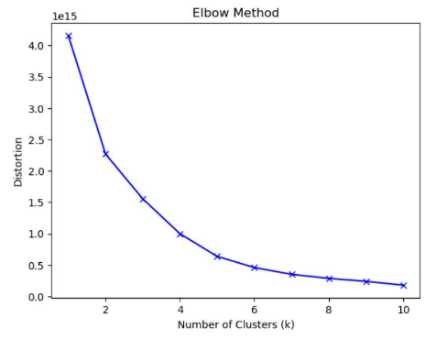
**Clustering:**

1. We loaded our data without the target column, which is the CNT\_CHILDREN column.

2. We preprocess the data by one-hot encoding categorical features, scales the numerical features, and then fits the KMeans model to the scaled data. The cluster centers are also printed to provide insight into the learned representations of the data.

3. We split the clustering into two different data files. We took the number of children and made it categorical in two forms. First file, according to a division we thought was reasonable (0, 1-3, 4-6, 7-10, 10+) according to the number of children suitable for a car. In this distribution, the variable Y is not equally distributed. We called this Unequal division into groups.  
In the second file, we divided the number of children a little more equally (0,1,2,3, and 4+). In this division, we deleted part of the data, because there were too many people without children. We called this Equal division into groups.

4.1 Visualize the elbow curve, which helps determine the optimal number of clusters for k-means clustering. The point indicates the number of clusters that capture the most significant amount of information without excessive complexity.



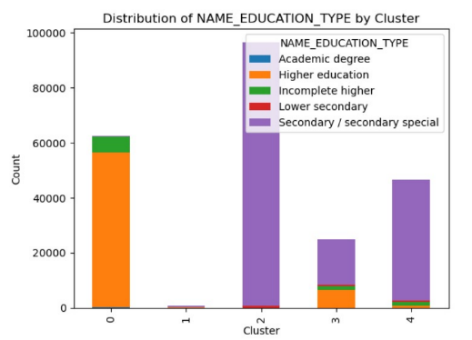
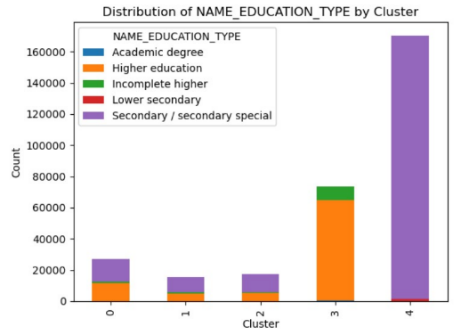
Unequal division into groups Equal division into groups

According to the graph we can see that we will divide into 5 clustering groups.

**Visualization:**

**Examples of the graphs we got:**

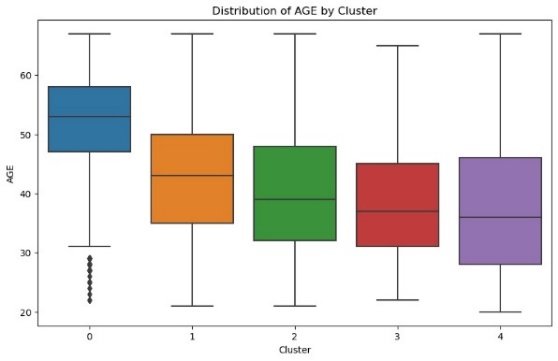
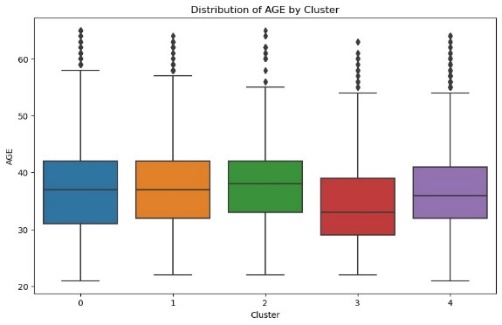
4.2 This visualization helps understand how the education types are distributed across different clusters, providing insights into the relationship between education and the clustering results.



Unequal division into groups Equal division into groups

In the graph where the groups are not equal, a person with a higher education is associated with groups 2, 3, 4, while in the graph where the division into groups is equal, a person with a higher education is associated with the 4th group.

4.3 This visualization helps understand the variation in age within each cluster and identify any differences or similarities in the age distribution across clusters.

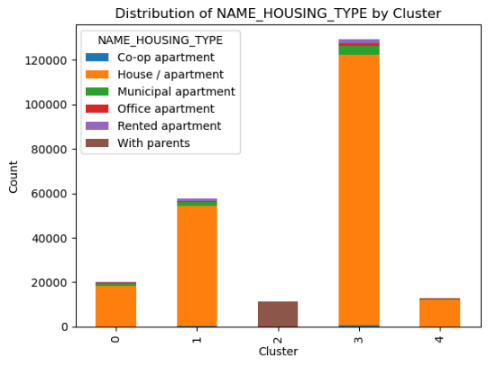


Unequal division into groups Equal division into groups

We can clearly see when the Y is divided in the form of categories that there is a difference between the ages. People in group 0 are more likely to be adults, around their 50’s.

On the other hand, in the second division there is no real difference between the ages in each group.

4.4 The model classifies well the group of people who live with the parents in the 2 cases In the case where the division into the number of children is more balanced, the model also classifies people who live in munificipal housing in a good way

 תמונה שמכילה טקסט, צילום מסך, תרשים, תוכנה

התיאור נוצר באופן אוטומטי

Unequal division into groups Equal division into groups

Additional graphs we made such as the distribution of marital status, the number of children, gender and the various flags did not show result Significance (some are attached in the presentation)

**Note**

We also tried to run density-based and hierarchy-based models, but since these models are heavy and our computing power is low, running the models crashed, so we decided to give up on running these models. In addition, we tried to build a K-means clustering model with a different distance function for each type of pitchers, but unfortunately this attempt failed

**Clustering Conclusions:**

There are no clear answers in the clustering model. We could not find interesting or relevant results in our models, even after trying two types of divisions.

**Division of tasks**

**Dvir -** Writing and editing the Milestone, Data Preprocessing, Model training, Visualizations, Writing the classification code, Writing the clustering code, Writing and editing Final Writeup.

**Shulamit –** Finding a data set for the project, Data Preprocessing, Model training, Exploratory Data Analysis, Visualizations, Writing the classification code, Writing the clustering code, Accuracy measures, Writing and editing Final Writeup.

**Yitzhak -** Writing and editing the project proposal, Exploratory Data Analysis, Writing the classification code, Writing and editing Final Writeup, Accuracy measures.