Overview

This report consists of 2 main sections. The first section is called “Task 1” which mainly contain analyses of distributions, comparison of these distributions, and statistical test. Each distribution analysis covers model 1, model 2, and the overall portfolio. In each analysis, I provided quantitative analyses (charts, tables) and qualitative conclusions. These qualitative results contain information about special concentrations of each distribution. In the future subsection of the mentioned section, I have presented comparisons of distributions to which these distributions belong to construction and validation datasets. I have used statistical components of the “Summary statistics” table to compare distributions with Data Spread, Variability, Central Tendency, etc. In the last subsection of “Task 1” section, I have implemented the Population Stability Index (PSI) to determine how much shift occurs over the period. Population Stability Index (PSI) is used for monitoring the stability of the model under Back Testing which I provide data shifts of distributions of important futures and model predictions.

As I mentioned we have the second section of the report which is called “Task 2”. In the mentioned part, I have created two Machine Learning models with the Scikit-Learn library. After the construction of the model, I selected one model according to different metrics. According to the requirements of the homework, I have generated 500 approved client dataset in which have been selected by the probability of approval according constructed model. I have provided tables, charts, and qualitative conclusions for model selection, 500 approved client determination, and finding essential futures for the model.

At the insudtry level, SPSS should be used for Task 2 and statistical test but my license expired therefore I have used Python-based solutions such as Pandas, Numpy, and Scikit-Learn. Scripts for the ML model will be provided in the solution folder.

Task 1

Analyze distributions

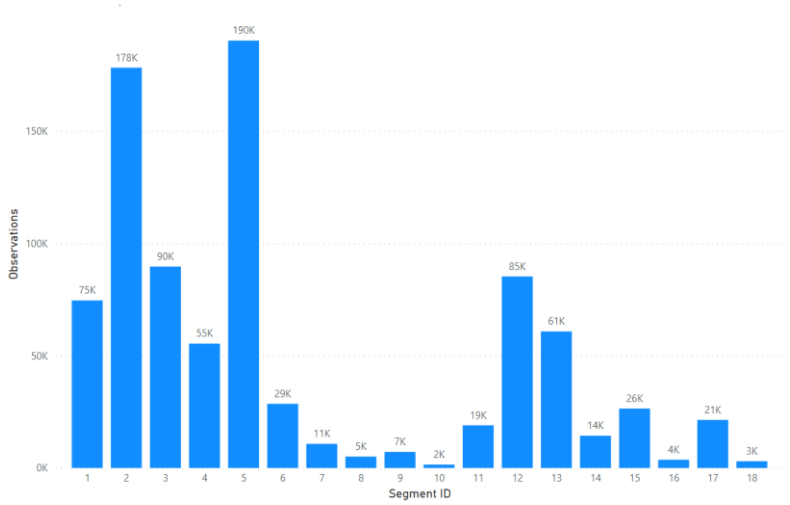
Construction

Distributions are analyzed at model and portfolio levels. In each analysis, special concentrations are highlighted. These concentrations can be represented as the lowest, and highest distribution points of given customer segments. Also, customer segments are highlighted if they share a total of 70 percent or 90 percent of the total distribution itself in a given segment range. There are null or zero distributions for some customer segments that typically belong to the validation dataset, they are highlighted in mentioned way. All qualitative conclusions are supported by charts and tables. Some quantitive analyses such as “Summary statistics” are used in the comparison of distributions and give us an overview of distribution with a given segment range.

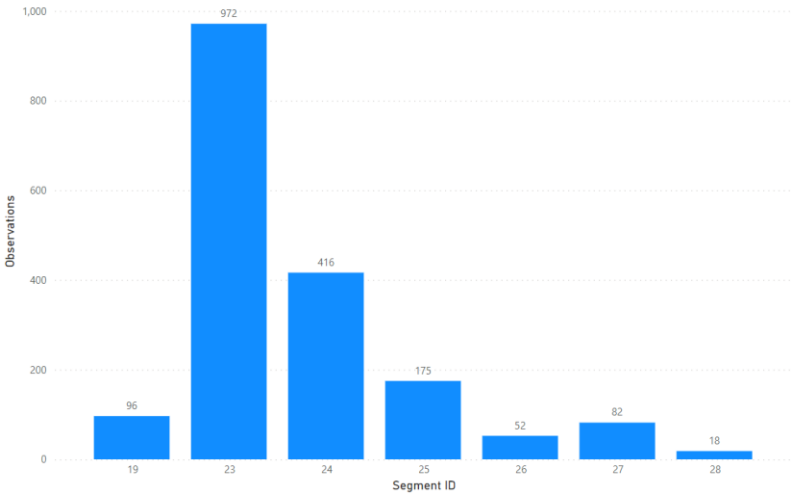
Distributions of observations

I would like to start analyzing the distribution of observations over segments which is used to train model 1. Observations are representations of financial services that our customers can use. For this reason, the distribution of observations represents the density of data over different consumed financial instruments. We can see special concentrations between the 1st and 7th customer segments which

contain more than 70 percent of total distribution. Also, we can see that 20 percent of the distribution is shared between the 11th and 18th. The lowest observation number belongs to the 10th customer segment. Opposite to the lowest point, the 5th user segment keeps most of the observations in the model 1 dataset.



The presented chart belongs to model 2. We can there is no equal distribution over customer segments in a dataset of model 2. Also, the 23rd user segment contains nearly 50 percent of the distribution. In other words, if we look at distributions with 23rd and 24th customer segments, we can easily notice two customer segments keep 90 percent of observations distribution. We had the lowest points in this distribution which can be presented 28th and 26th client segments respectively.

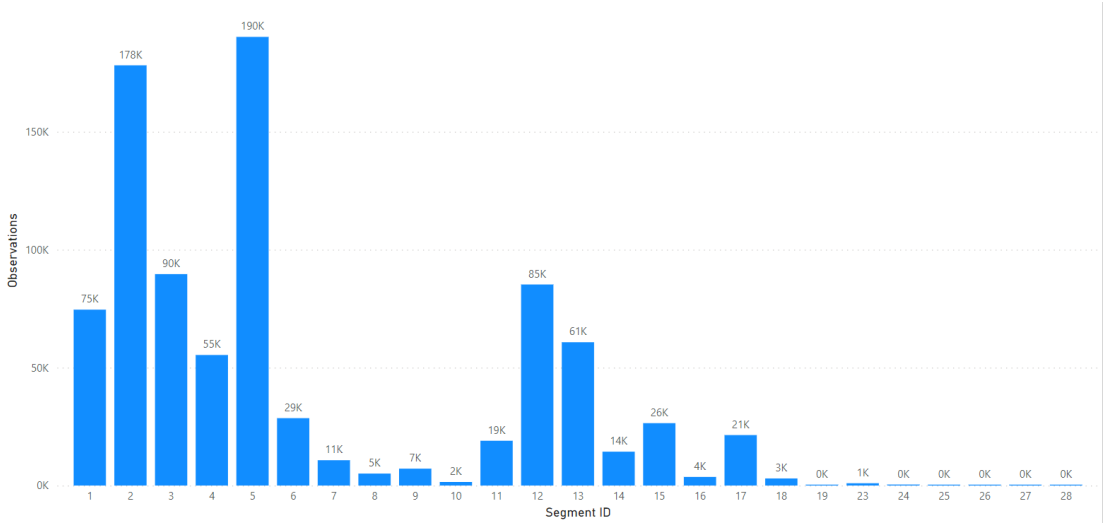


We have analyzed the observation distribution of two models but we need to analyze the big picture now. As we can see, most of the observations belong to model 1, and little part of consumed financial instruments are represented in model 2. Also, we can see observations between the 1st and 5th client segment keep more than 70 percent of total observations in total.

Summary statistics	
Mean	35067
Median	10728

Mode	190238
STD	51990.35

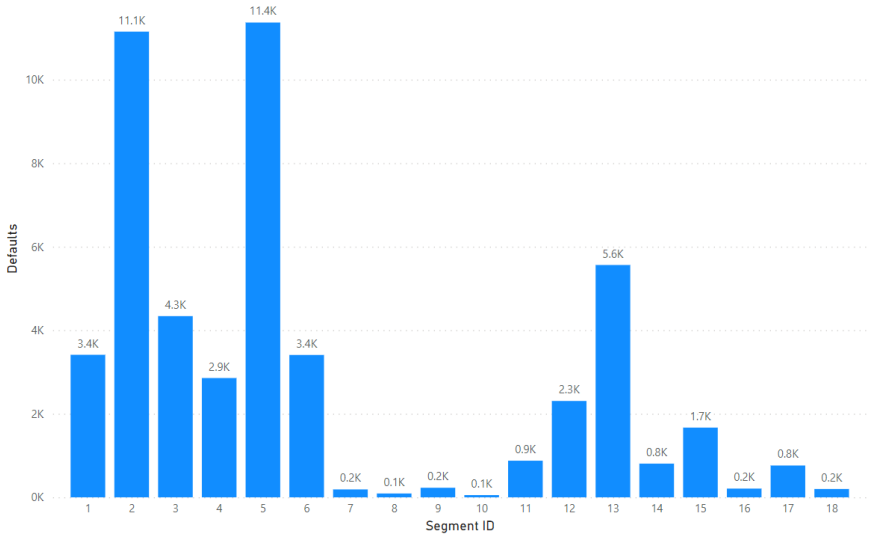
Table 1.



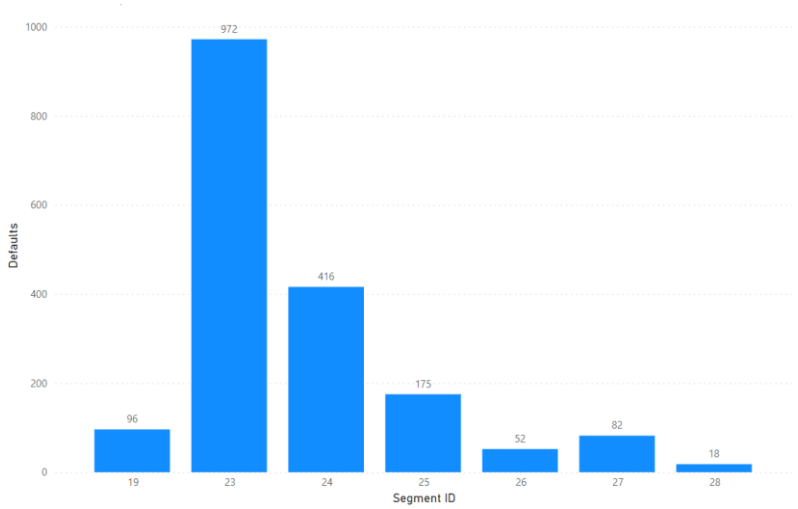
Distributions of defaults

We can distribute defaults over customer segments. We can probably same percentage of distribution as the distribution of observations. Because more observation of a given customer segment means more risk in non-payment for consumed financial instruments. Defaults are important for the prediction of CCF and in related credit risk models therefore analyzing the distribution of defaults is important to model 1.

We can see that most distribution observation user segments are located between 1 and 6 which saves more than 90 percent of total distribution. Also, we can see significant distribution varies between the 11th and 18th customer segments. Two customer segments share the same highest disturbed points which are the 5th and 2nd client segments respectively. Also, if we look at the lowest points they are 8th, 10th, and 7th respectively. The defaults are not distributed equally over customer segments because observations are not distributed equally over given customer segments. It can lead to some overfitting in the credit risk model for the prediction of CCF.



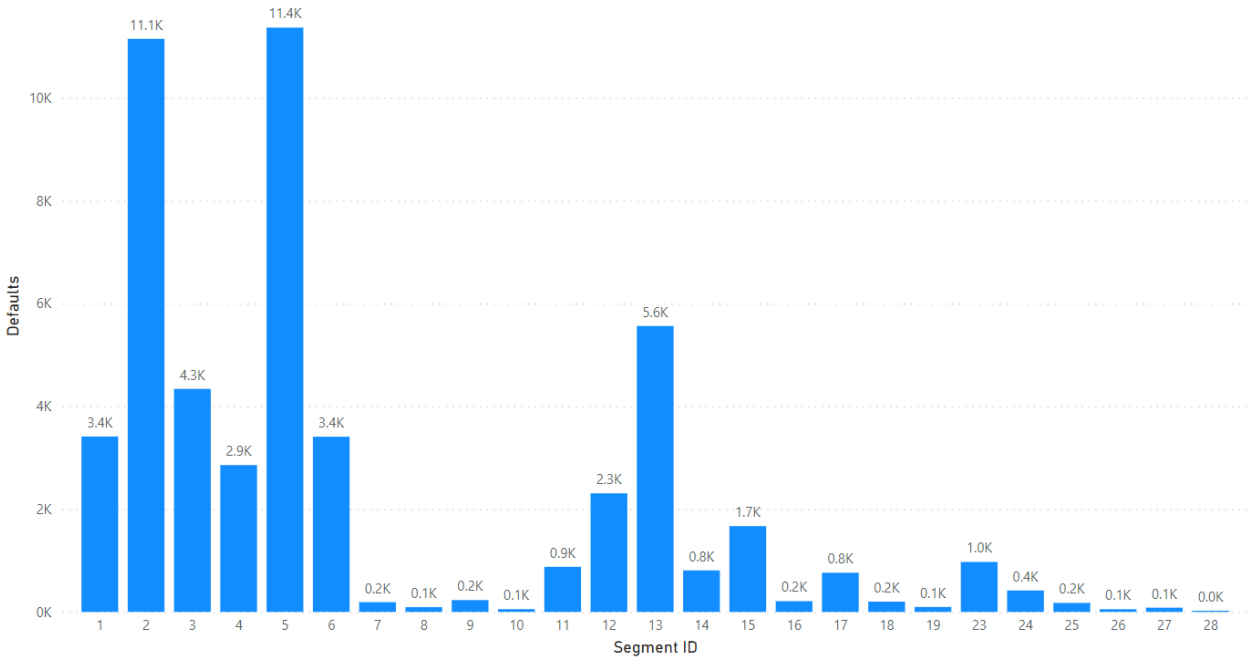
For the nature of given data, the distribution of defaults are same as the distribution of observation. Therefore, highlighted concentrations for observations in model 2 can be applied for defaults. As I mentioned this imbalance in the dataset can lead to overfitting in the model.



Most of the concentrations in distribution are mentioned early but we had a chance to look big picture through two models together. As I mentioned in the distribution of observations, model 2 is used as a small part of the distribution. This small part of the distribution of default can lead to underfitting in validation. We will use descriptive information from the summary statistics table to compare the construction and validation distributions of defaults. Also, summary statistics can tell us that distributions are concentrated in the same customer segments.

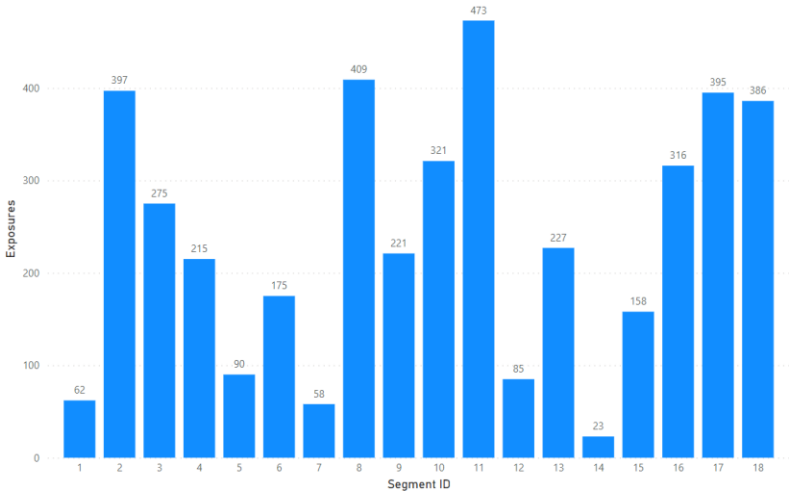
Summary statistics	
Mean	2052
Median	762
Mode	11367
STD	3104.09

Table 2.



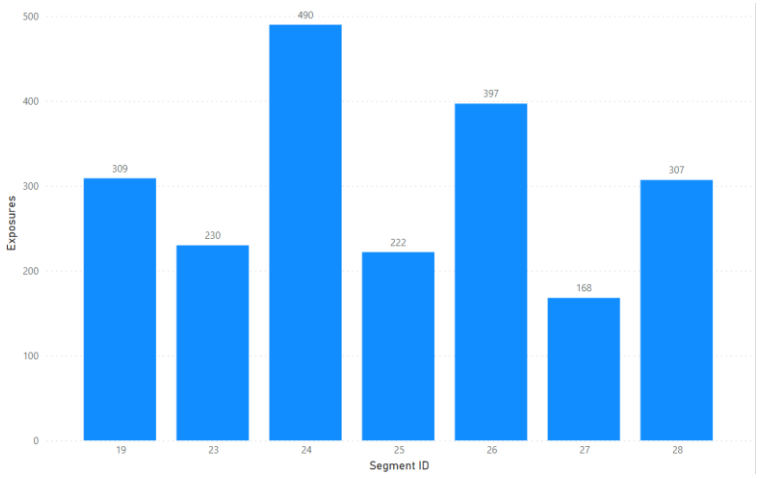
Distributions of exposures

Exposures are important variables for construction models that represent a potential loss in some form of financial instrument. For this reason, it can be considered a real money representation of losses. We can see a slightly equal distribution compared to distribution observations and defaults across client segments. However, we can still see some concentrations in distribution. We have a customer segment that is most likely to cause us to lose money, this customer segments are 11th with 473 million euro. Also, we have a client segment that can just 23 million lost if there is some default on a loan or consumed financial instrument. This distribution shows us to model one can be protected from overfitting.



We can same imbalanced distribution across different user segments which these segments are used for construction model 2. However, we have the lowest and highest distribution probations for some customer segments.24th customer segment can result in 490 million euros lost if default occurs on

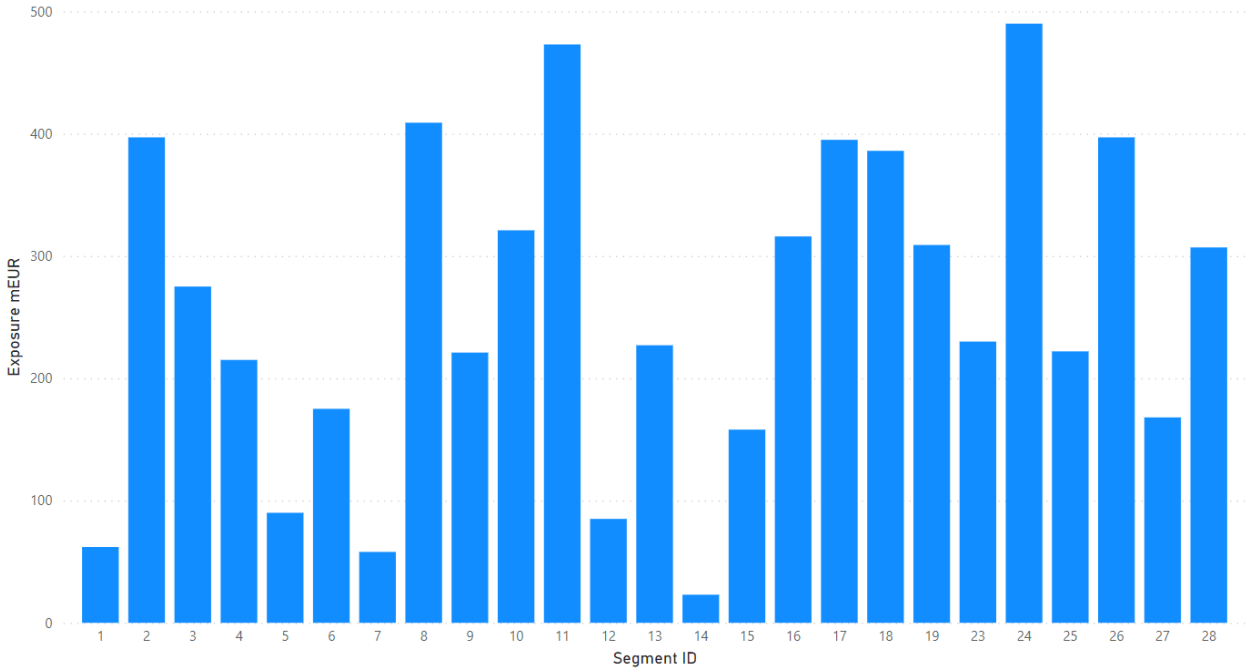
customer loans or credits. Also, the safest customer group in this distribution is 27th which holds a relatively small amount of exposure amount.



There are not many more significant concentrations through distributions of exposure across customer segments.

Summary statistics	
Mean	256
Median	230
Mode	490
STD	131.49

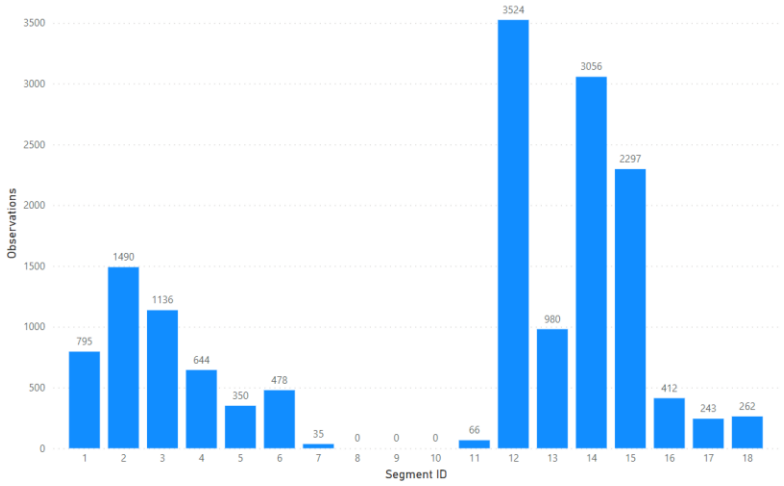
Table 3.



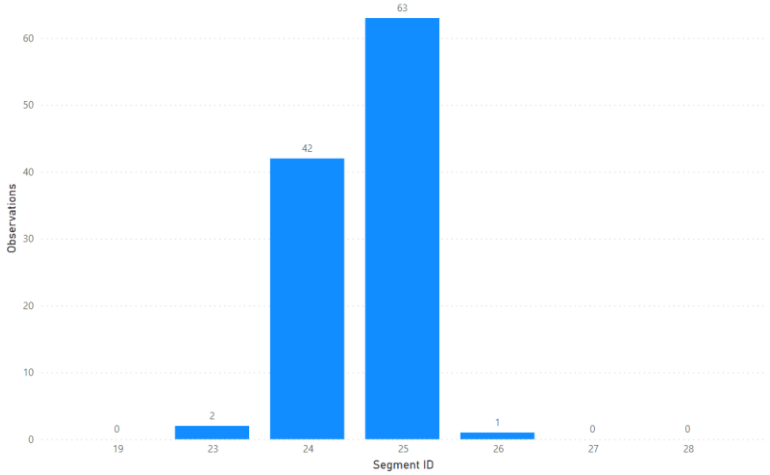
Validation (2020)

Distributions of observations

We can see concentrations occur on the right side of the histogram which is opposite to the construction distribution of observations. Also, we can see there is some elimination of customer segments by 2020. Eliminating client segments are represented in 8th, 9th, and 10th IDs. Distribution of customer segments between 12 and 15 share more than 70 percent total population of observations. Also, The highest distribution rates are represented on the left side of the chart. The 12th customer segment can be seen with the highest distribution rate. This observation disturbance difference between the construction and validation dataset can affect defaults which has a direct effect on the performance of the model. We can overfit and prediction shifts over the period for this reason.



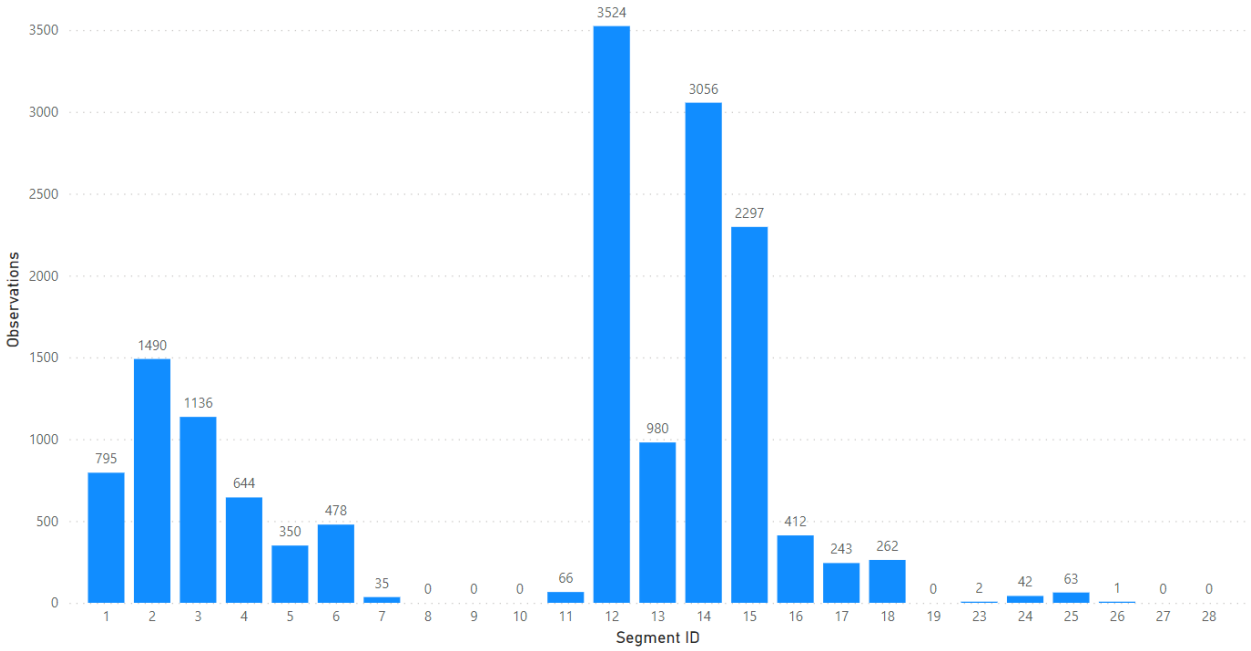
We can see some eliminated or unsuspected customer segments which are the 19th, 27th, and 28th client segments. We can see the highest concentration of observations belongs to the 24th and 25th customer segments. Also, the lowest customer segments are 26 and 23 respectively.



According bird's view of distribution in the Clustored Bar Chart, we can say 12th, 14th, and 15th customer segments share more than 80 percent of the total observation distribution in total. Eliminated customer segments can have some effect decrease in the performance of model 2 and model 1.

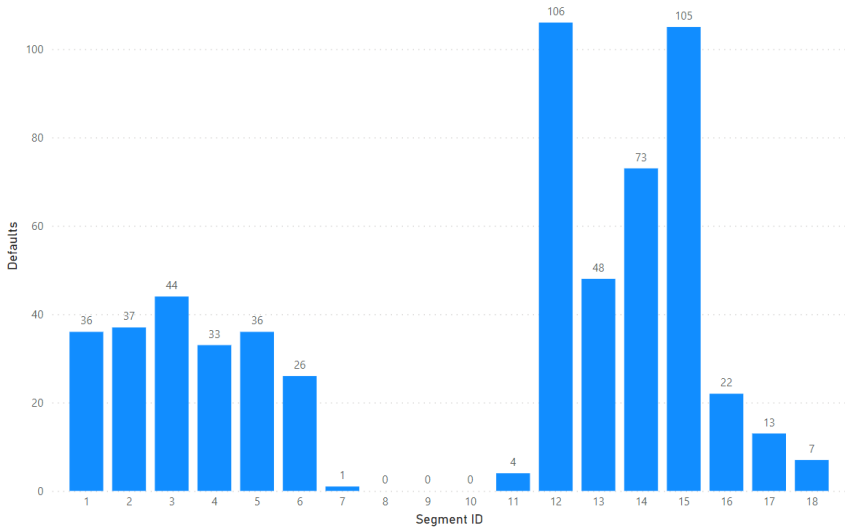
Summary statistics	
Mean	635
Median	243
Mode	3524
STD	961.86

Table 4.

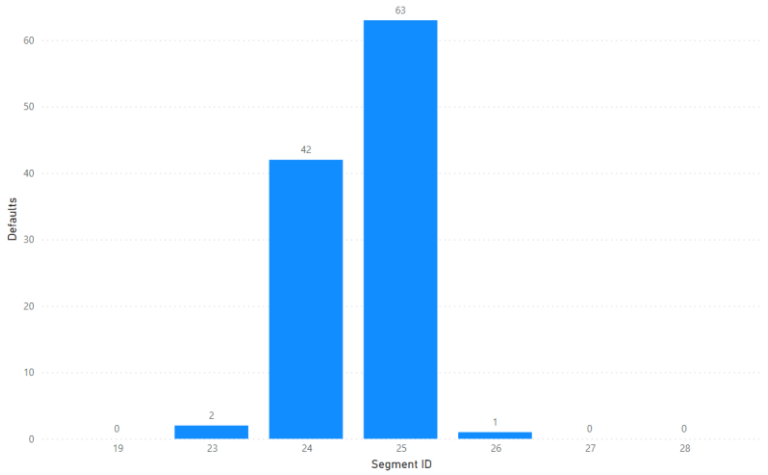


Distributions of defaults

We can see the same tendency in defaults which have the same disturbance rates as observations. It may happen. We can see the same lack of distribution for the 8th, 9th, and 10th customer segments. The distribution between 12 and 15 contains more than 70 percent of the distribution of defaults for model 1. The right side of distribution from the 1st and 6th customer segments can belong to equal distribution but we can see the more viewed distribution for the left side. Also, the two customer segments share the highest distribution of defaults, they are the 12th and 15th segments respectively. We have the lowest distribution rate for the 7th customer segment which is just 1 default in the validation dataset.



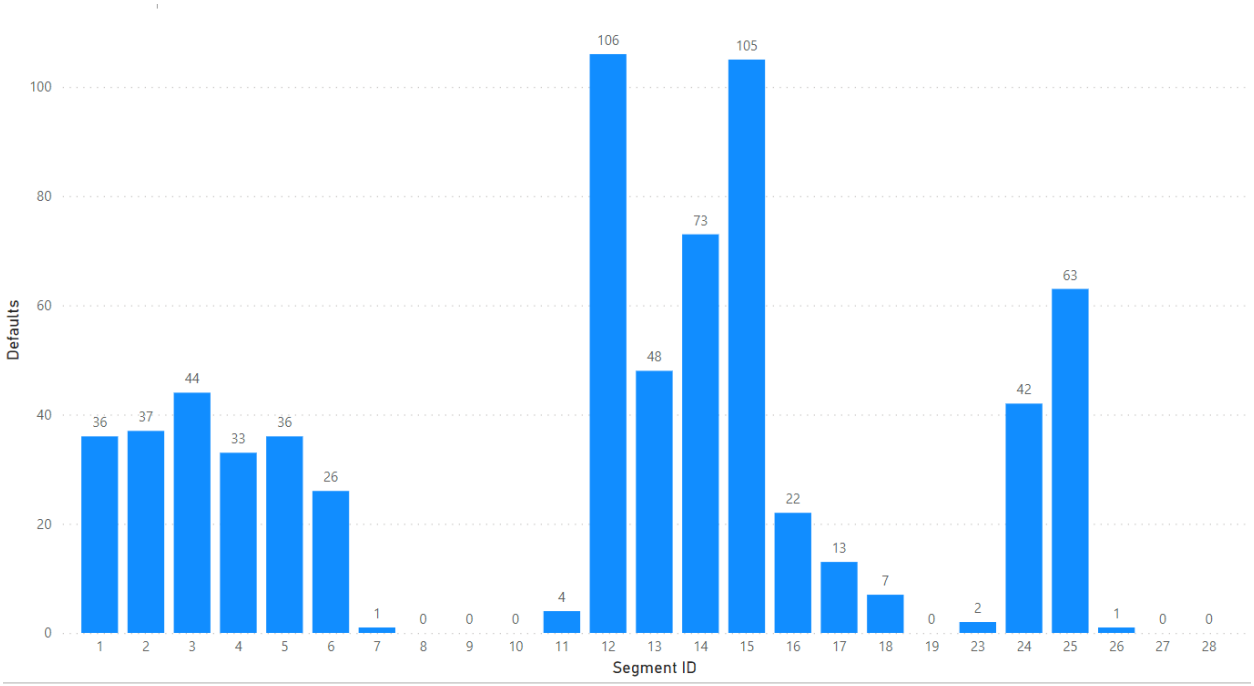
We can see the same rate of disruption of defaults with the distribution of observations. Also, we have some eliminated or missing customer segments which is not used for validation of model 2. These missing values can make inappropriate model validation for model 2 and it can affect predictions of CCF.



I would like to start with the highest distribution points which are between 12 and 15 contain more than 80 percent of the total distribution. Also, we can see summary statistics which can provide qualitative analyses.

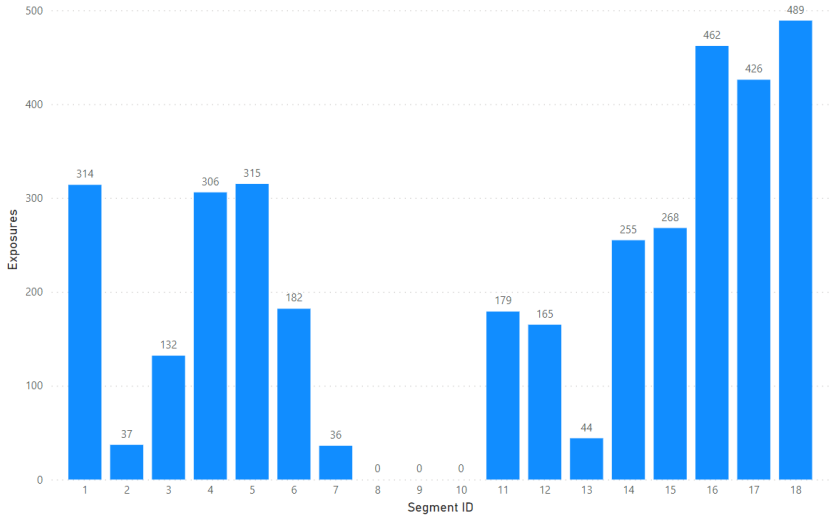
Summary statistics	
Mean	28
Median	22
Mode	106
STD	31.29

Table 5.



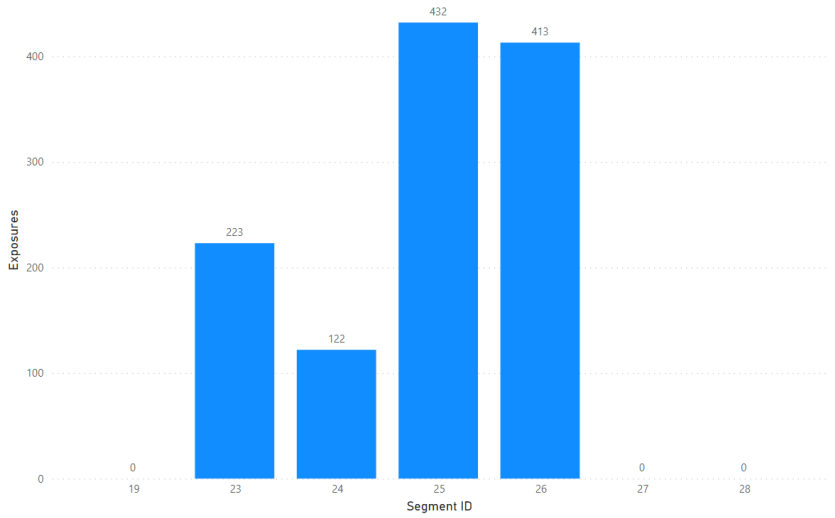
Distributions of exposure

I would like to give observation distribution details over segments. But unfortunately, we can see some missing exposures for the 8th, 9th, and 10th customer segments. Also, distributions between the 14th and 18th customer segments contain more than 80 percent of the total distribution. Following them, we have the lowest and highest distribution points which are the 7th and 18th customer segments respectively. Exposure real impact if there are some defaults in consumed financial instruments such as loans therefore, we can see some similar distribution rates with defaults.



Model 2 has the same missing or eliminated segments as defaults. But we can see the 26th and 23rd segments belong to more money lost if there are some defaults. It is interesting because default distribution is less in these segments for some reason. It may happen in terms of consumed financial

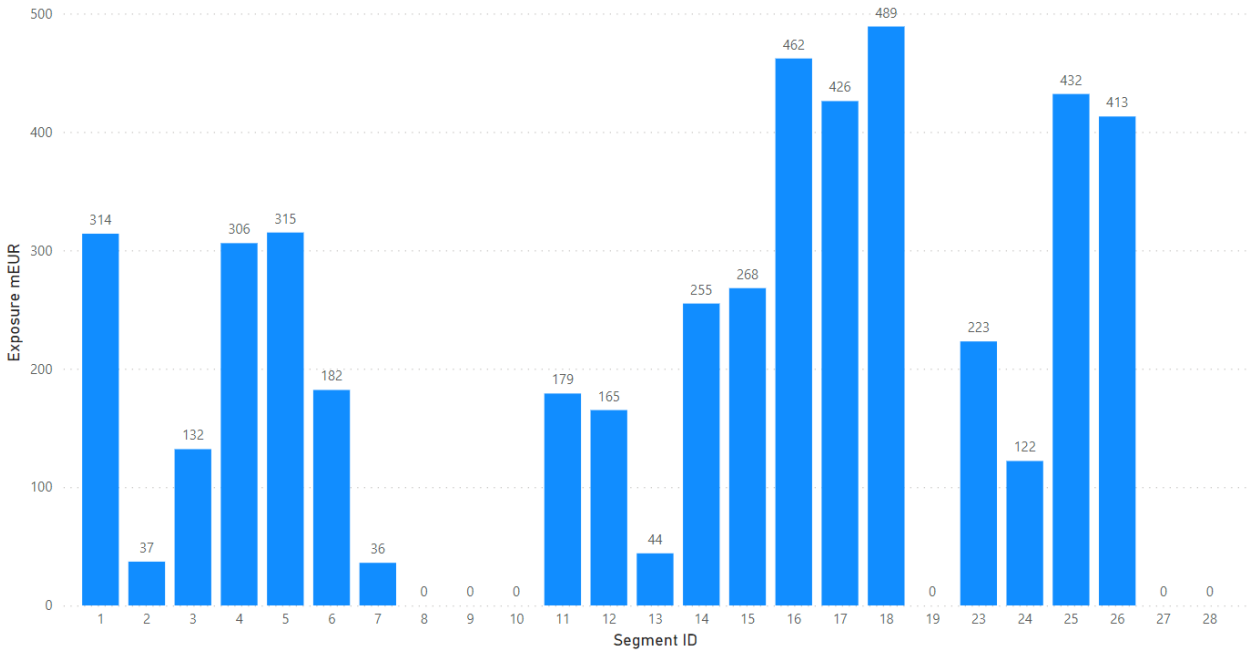
instruments. Also, we can notice, that the highest distributions are shared by the 25th and 26th customer segments respectively.



Exposure distribution across segments is continuous except at some points such as 8, 9, 10, 19, 27, and 28. Also, the highest distributions of exposure are shown between the 16th and 18th customer segments which keep more than 50 percent of the total distribution. As we can notice from the summary statistics table, it is acceptable for models.

Summary statistics	
Mean	192
Median	179
Mode	489
STD	164.56

Table 6.



Comparing distributions

Comparing distributions over segments is important because we can understand anomalies through models and total portfolios. I would like to present these comparisons qualitative conclusion by comparing summary statistics of portfolios. I mainly focus on the statistical components of each distribution in my comparisons.

Observations

Statistics components of Table 1 and Table 4 are used for these comparisons. The construction dataset has a significantly higher mean number of observations per Segment ID (35067) compared to the validation dataset (635), indicating that, on average, segments in the construction dataset have more observations. The construction dataset has a much larger standard deviation (51990.35) compared to the validation dataset (961.86). This suggests that the number of observations per Segment ID varies widely in the construction dataset, indicating a more spread-out distribution, while the validation dataset has a more consistent distribution with lower variability. The median observation count in the construction dataset is 10728, which is significantly higher than the median in the validation dataset (243). This means that half of the segments in the construction dataset have 10728 or more observations, while half of the segments in the validation dataset have 243 or fewer observations. The mode of 190238 in the construction dataset indicates that there is a specific Segment ID with an exceptionally high number of observations. In contrast, the mode of 3524 in the validation dataset suggests a different Segment ID with the highest frequency of observations, but the count is significantly lower than that in the construction dataset. The construction dataset has a wider range of observation counts, as indicated by the higher standard deviation and the presence of a much larger mode. This suggests that the construction dataset contains segments with both very high and very low numbers of observations. On the other hand, the validation dataset has a narrower range, with observation counts clustered around the mode.

In summary, the construction dataset exhibits a broader and more varied distribution of observations across Segment IDs, including segments with extremely high observation counts. In contrast, the validation dataset has a narrower and more consistent distribution, with fewer extreme values.

Defaults

Statistics components of Table 2 and Table 5 are used. The construction dataset has a significantly higher mean number of defaults per Segment ID (2052) compared to the validation dataset (28). This indicates that, on average, segments in the construction dataset experience a much higher number of defaults. The construction dataset has a larger standard deviation (3104.09) in defaults compared to the validation dataset (31.29). This suggests that the number of defaults per Segment ID varies widely in the construction dataset, indicating a more spread-out distribution, while the validation dataset has a more consistent distribution with lower variability. The median default count in the construction dataset is 762, which is significantly higher than the median in the validation dataset (22). This means that half of the segments in the construction dataset have 762 or more defaults, while half of the segments in the validation dataset have 22 or fewer defaults. The mode of 11367 in the construction dataset indicates that there is a specific Segment ID with an exceptionally high number of defaults. In contrast, the mode of 106 in the validation dataset suggests a different Segment ID with the highest frequency of defaults, but the count is significantly lower than that in the construction dataset. The construction dataset has a wider range of default counts, as indicated by the higher standard deviation and the presence of a much

larger mode. This suggests that the construction dataset contains segments with both very high and very low numbers of defaults. On the other hand, the validation dataset has a narrower range, with default counts clustered around the mode.

In summary, the construction dataset exhibits a broader and more varied distribution of defaults across Segment IDs, including segments with extremely high default counts. In contrast, the validation dataset has a narrower and more consistent distribution, with fewer extreme values, indicating a substantially lower default rate for the segments in the validation dataset.

Exposure

Statistics components of Table 3 and Table 6 are used. The construction dataset has a higher mean Exposure mEUR per Segment ID (256) compared to the validation dataset (192). This indicates that, on average, segments in the construction dataset have a higher exposure in million Euros (mEUR). The construction dataset has a smaller standard deviation (131.49) in Exposure mEUR compared to the validation dataset (164.56). This suggests that the Exposure mEUR per Segment ID varies less in the construction dataset, indicating a more concentrated distribution, while the validation dataset has a wider spread with higher variability. The median Exposure mEUR in the construction dataset is 230, which is higher than the median in the validation dataset (179). This means that half of the segments in the construction dataset have an exposure of 230 mEUR or more, while half of the segments in the validation dataset have 179 mEUR or less. The mode of 490 in the construction dataset indicates that there is a specific Segment ID with the highest frequency of a 490 mEUR exposure. In contrast, the mode of 489 in the validation dataset suggests a different Segment ID with the highest frequency of exposure, but the exposure level is very similar to that in the construction dataset. The construction dataset has a narrower range of exposure values, as indicated by the smaller standard deviation and the presence of a mode close to the mean. This suggests that the construction dataset contains segments with more consistent exposure levels. In contrast, the validation dataset has a wider range, with exposure levels distributed over a larger span.

In summary, the construction dataset exhibits a higher average exposure level and a more concentrated distribution of Exposure mEUR across Segment IDs. On the other hand, the validation dataset has a lower average exposure level, higher variability, and a wider spread of exposure values, indicating a different risk profile or exposure pattern for the segments in the validation dataset compared to the construction dataset.

Statistical test

I intend to perform Back Testing which is an important part of the Risk component of the Validation of the Rating System. There are various levels of backtesting, but for this task, I will focus on Level 0 which ensures the stability of our model validation process, allowing us to assess how much the distribution shifts between different periods. The primary metric we will employ for this assessment is the Population Stability Index (PSI) which is used in Level 0 of Back Testing

In our dataset, we have columns for Observations, Defaults, CCF (Credit Conversion Factor), and Exposure mEUR. However, for the purpose of calculating the PSI-based valuation metrics, I will only utilize Defaults, CCF, and Exposure mEUR. Observations indicate the number of financial instruments or accounts

within each segment but do not significantly impact the prediction model. Hence, we will focus solely on the three distributions across segments to calculate our metrics.

The results will be presented through quantitative analyses, such as tables, and qualitative methods. I will conduct statistical tests at both the model and portfolio levels, excluding null and zero values within each segment.

The test result of defaults

The results of the PSI comparison table show there are dramatic changes through the years and the actual value has a big amount of gap. It has a positive impact on credit risk assessment because customers can pay their debts more than usual with actual values of 2020. This comparison belongs to model 1.

Segments	PSI comparison
1	153.66
2	634.294
3	197.137
4	125.925
5	652.093
6	164.899
7	9.854
11	47.057
12	67.797
13	262.054
14	17.637
15	43.261
16	4.21
17	30.492
18	6.427
PSI Actual	2416.797

There are some PSI calculated measures for model 2. Comparison table does not show dramatic differences over a year but it is still high for PSI measurement. We can understand the customer in selected segments is more responsible for their loans.

Segments	PSI comparison
23	60.006
24	8.576
25	1.144
26	2.015
PSI Actual	**71.741**

The test result of CCF

There is an interesting situation for us in this table in which we can see there are no significant changes over the period but the model faced big value shifts. We can understand we do not need to keep more money for risk financial instruments but the model should be changed according to PSI value. We have some CCF predictions which show we should not use the model for the 11th and 15th customer segments. If we can find and solve issues for this user segment, we can keep model 1 for a bit longer with small maintenance.

Segments	PSI comparison
1	0.154
2	0.095
3	0.115
4	0
5	0.058
6	0.018
7	0.007
11	0.233
12	0.012
13	0.012
14	0.071
15	0.387
16	0.021
17	0.002
18	0.153
PSI Actual	1.338

Model 2 faces some change rates over the years which means less money is needed for each loan or financial instrument in total for some individual customer segments. Model 2 should not be used for a lot because PSI value alarm on this. It shows there are some macroeconomic changes in the table.

Segments	PSI comparison
23	0.097
24	0.44
25	0.344
26	0.47
PSI Actual	**1.351**

The test result of exposure

According to model 1, we can see there are more shifted values over time for exposure money. It means actual risked money is decreased in validation. This is not the same for some selected segments as we can see 6th and 7th customer segments.

Segments	PSI comparison
1	4.088
2	8.543
3	1.05
4	0.321
5	2.819
6	0.003
7	0.105
11	2.857
12	0.531
13	3.003
14	5.581
15	0.581
16	0.555
17	0.023
18	0.244
PSI Actual	30.304

We faced the same situation in model 2 we have some customer segments which has more financial instruments on risk. 23rd and 26th show we have more money risk for these user segments.

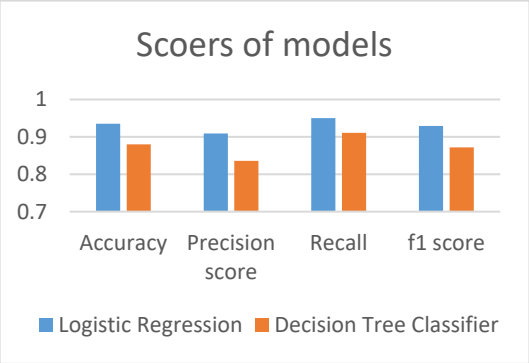
Segments	PSI comparison
23	0.002
24	5.117
25	1.398
26	0.006
PSI Actual	**6.523**

Task 2

The solution for the second task mainly contains a constructed Machine Learning model and prediction of the probability of each new input by this model. Also, I have answered extra questions with relevant details and visualizations. Most answers and qualitative results contain information about techniques and methodology that I have used. At the industry level, SPSS can be used for this purpose, but my license is expired therefore, I have used a Python-based Machine Learning library which is called Scikit-Learn and Pandas library for manipulation of data. All code blocks and result datasets can be found in the solution folder.

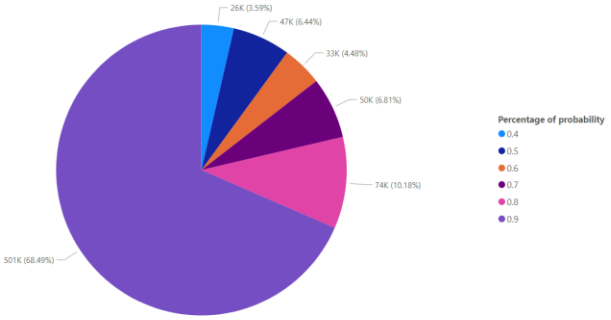
I would like to support predictive results with quantitative and qualitative analyses. At first glance, I would like to start model construction in which the “Applications_Decided_On” sheet in the second

dataset is used for training, and all categorical data is converted. I have constructed two models and we can see the compassion of their scores.



Model	Accuracy	Precision	Recall	f1 score
Logistic Regression	0.935	0.909	0.95	0.929
Decision Tree Classifier	0.88	0.836	0.911	0.872

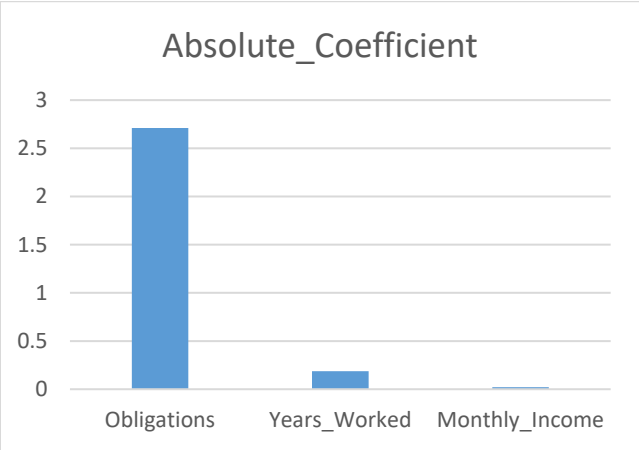
As we can notice Logistic Regression can get enough good results in different scores. For this reason, I decided to implement Logistic Regression for the prediction of approvals probabilities. The final dataset can be found under with probability of values. Clients who have a probability of more than 50 percent, can get positive approval. And we can see nearly 90 percent of 500 clients have a possibility percentage more than 50 percent. But we can notice small and lucky customers who can get approval because they are placed last places in the 500 quarter of loans. These lucky customers are presented in 19 total.



Number of clients	Percentage of probability
339	90%
52	80%
36	70%
24	60%
30	50%
19	40%
Total	**500**

As I mentioned, there are 2 extra questions, I would like to answer the first question. If we put 200 Clinet_ID data into a model, we get a 30 percent probability of getting approval which means there is a low probability.

As we notice from the coefficient table, we can understand “Obligations” column has importance in prediction. The coefficient can be calculated by the Scikit-Learn library and we can see its distribution across futures in the Clustered Column Chart.



Feature	Absolute_Coefficient	Coefficient
Obligations	2.711048	-2.711048
Years_Worked	0.186424	0.186424
Monthly_Income	0.021207	0.021207