Nachvollziehbare Schritte
Churn modelling prediction

Sanja Srdanovic, 18.03.2022

1. Kurze Darstellung des Problembereichs / Aufriss des Themas

1.1 Inhaltlich

In this project, a *Workflow* in *KNIME* is created to examine a data set on *Churn Modelling* from *Kaggle* https://www.kaggle.com/shrutimechlearn/churn-modelling. The goal is to predict whether bank customers will churn or not, based on various characteristics such as customers' age, gender, salary, credit score, balance, number of products etc. Additionally, a connection between churn and some characteristics of customers is visualised in *Tableau*.

1.2 Begründung des Themas

Analysing and forecasting whether customers will churn or not is very important for banks and some companies, who would of course like to keep their customers. This can be predicted based on several characteristics and it is significant to find out which of them are affecting whether members have exited or not, so that the bank can improve their services or optimize some features to prevent people from churning, or at least to minimize the number of members who exit.

2. Nachvollziehbare Schritte

2.1 Der Stand der Forschung / Auswertung der vorhandenen Literatur / Tutorials ...

There are various approaches to *Churn model prediction*. In my project, I created a *Workflow* in *KNIME* and for the analysis, I applied the models *Decision Tree, Random Forest, Logistic Regression* and *XGBClassifier* in order to check which of them would yield the best accuracy in predicting whether the customers will leave the bank or not. For the analysis in *KNIME*, I used *Tutorials 11–13* as a guide, for the visualization in *Tableau Tutorial 14*.

2.2 Fragestellung

1. How well can churn be predicted and which model gives the best accuracy?
2. Which factors have the most influence on customers' churn?

2.3 Methode - KNIME

2.3.1 Loading the dataset

The dataset was downloaded as a CSV file and loaded into *KNIME* using the *File Reader* node. The column headings were taken over from the table.

2.3.2 Inspecting the dataset

In order to get the first overview of the dataset, I used the following nodes: *Extract Table Dimensions*, *Extract Table Spec, Statistics, Boxplot* and *Linear Correlation*. From the *Table Dimensions*, it can be observed that there are 10 000 rows and 14 columns. The columns are the following: *RowNumber, CustomerId, Surname, CreditScore, Geography, Gender, Age, Tenure, Balance, NumofProducts, HasCrCard, IsActiveMember, Estimated Salary* and *Exited*. The last column is our output value, i.e., it tells us whether a customer has churned or not (1 = yes, 0 = no). The columns are of different types, including both continuous and categorical variables. The box plots give us an overview of what continuous variables look like and the correlation heatmap shows a correlation between the variables. What can also be observed is that there are no zero, i.e., no missing values in the dataset.

2.3.3 Pre-processing

In order to prepare the data for the analysis, I did a couple of things. There were no missing values, so I could proceed with other pre-processing tasks. First, I removed the column *Row Numbers* and then I converted the column *Exited* from *Number to String*. I normalized the data using *Min Max Scaler* in the *Normalizer* node and I checked the results after the normalisation with the *Box plot* again. The values are now between 0 and 1, there are some outliers, but the machine learning models can deal with them very well. Additionally, with the *Numeric Binner* node, I created a few new columns, where I divided some columns into range groups – *Age, Credit Score, Salary groups* and *Balance*. Then I saved this Table with a CSV writer, so that I could use it for the visualisations in *Tableau*.

2.3.4 Splitting the data into train and test

The data were prepared for analysis by nodes *Color Manager* and *Partitioning*. In the *Color Manager* it was configured that churn is red (1) and no-churn is blue (0). In the *Partitioning* node, the data are split into train and test data by using stratified sampling: 80% for the training data that would be used for the learning phase (8000 entries), and the rest 20% for the test data for validation (2000 entries).

2.3.5 Analyses

Given that we are dealing with the classification problem, the methods I used for the analysis are *Decision Tree, Random Forest, Logistic Regression* and *XGBClassifier.* The respective models are trained with the training data. For evaluation, the churn is predicted with the test data and the accuracy is calculated by the *Scorer (JavaScript).*

2.4 Methode - Tableau

In *Tableau*, I loaded the *CSV Table* created in *KNIME*. I wanted to display the number of countries where the bank is present, how many customers there are, how many male and female customers, the number of active members, how many have exited and how many members have a credit card. Then I

created a histogram of how age is distributed among the number of customers and then how many of them exited based on their age. Also, I created a bar chart for each country to examine how many customers there are in each country, and how many of them left the bank. Additionally, I created a map, where some of the important information about churn for each country are displayed as details. Finally, I created a sheet called *Parameter* where the variables *Gender* and *Exited* are rows, and *calculated measures* as columns. The parameter *Measures* includes *Age*, *Balance*, *Credit Score*, *Salary*, *Tenure* and *Number of products*, thus we can observe how many male and female customers exited based on these features. All the sheets were then combined into a *Dashboard*. I used a couple of horizontal layers to get the layout I wanted and did some formatting. Additionally, I created several filters for a better and more interactive display of the results. Moreover, I created a *Story* on churning based on the different groups of features I created earlier in *KNIME*. The screenshot will be shown in the Results Section.

2.5 Ergebnisse

2.5.1 KNIME

The models used for the analysis in *KNIME* were *Decision Tree, Random Forest, Logistic Regression* and *XGBoost*. The results are compiled in the *Scorer (JavaScript)* as a confusion matrix and overall statistics including accuracy.

Decision Tree

Confusion Matrix

	0 (Predicted)	1 (Predicted)	
0 (Actual)	1433	160	89.96%
1 (Actual)	195	212	52.09%
	88.02%	56.99%	

Overall Statistics

Overall Accuracy	Overall Error	Cohen's kappa (κ)	Correctly Classified	Incorrectly Classified
82.25%	17.75%	0.434	1645	355

Random Forest

Confusion Matrix

	0 (Predicted)	1 (Predicted)	
0 (Actual)	1537	56	96.48%
1 (Actual)	216	191	46.93%
	87.68%	77.33%	

Overall Statistics

Overall Accuracy	Overall Error	Cohen's kappa (κ)	Correctly Classified	Incorrectly Classified
86.40%	13.60%	0.509	1728	272

Logistic Regression

Confusion Matrix

	0 (Predicted)	1 (Predicted)	
0 (Actual)	1519	74	95.35%
1 (Actual)	279	128	31.45%
	84.48%	63.37%	

Overall Statistics

Overall Accuracy	Overall Error	Cohen's kappa (κ)	Correctly Classified	Incorrectly Classified
82.35%	17.65%	0.330	1647	353

XGBoost

Confusion Matrix

	0 (Predicted)	1 (Predicted)	
0 (Actual)	1519	74	95.35%
1 (Actual)	279	128	31.45%
	84.48%	63.37%	

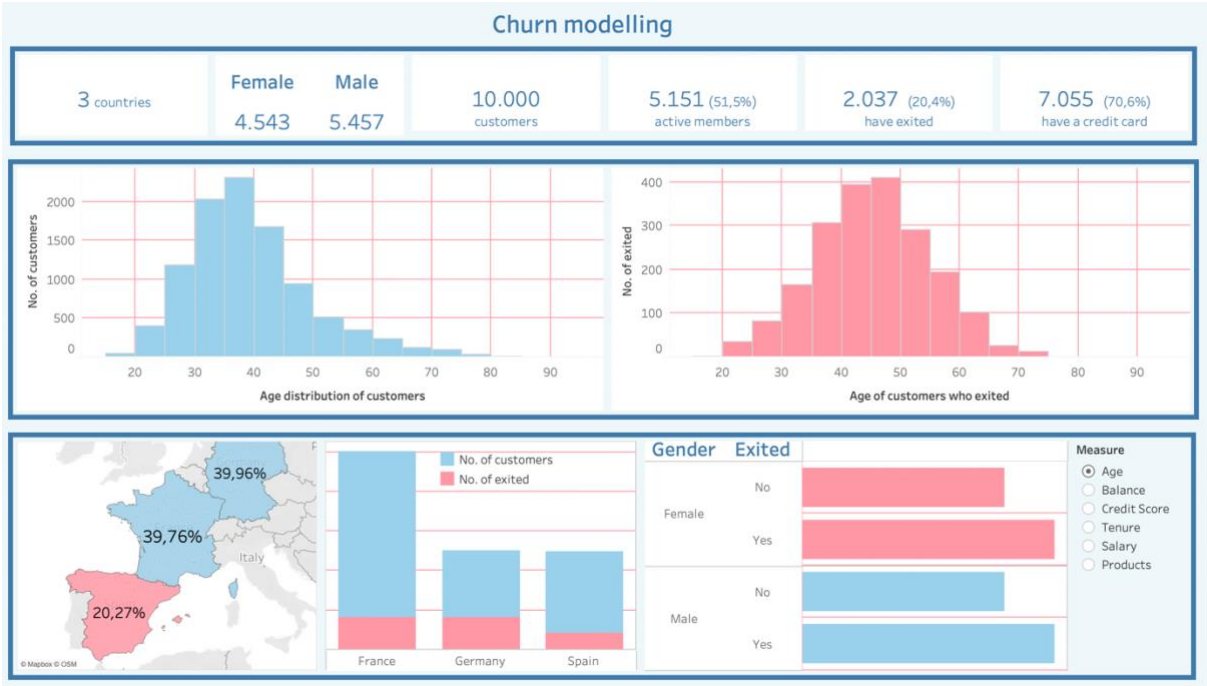
Overall Statistics

Overall Accuracy	Overall Error	Cohen's kappa (κ)	Correctly Classified	Incorrectly Classified
82.35%	17.65%	0.330	1647	353

The results indicate that the model that best performed is *Random Forest*, with an accuracy of **86,40 %**, all the other models have similar accuracy, a little bit lower than *Random Forest*, namely 82 %.

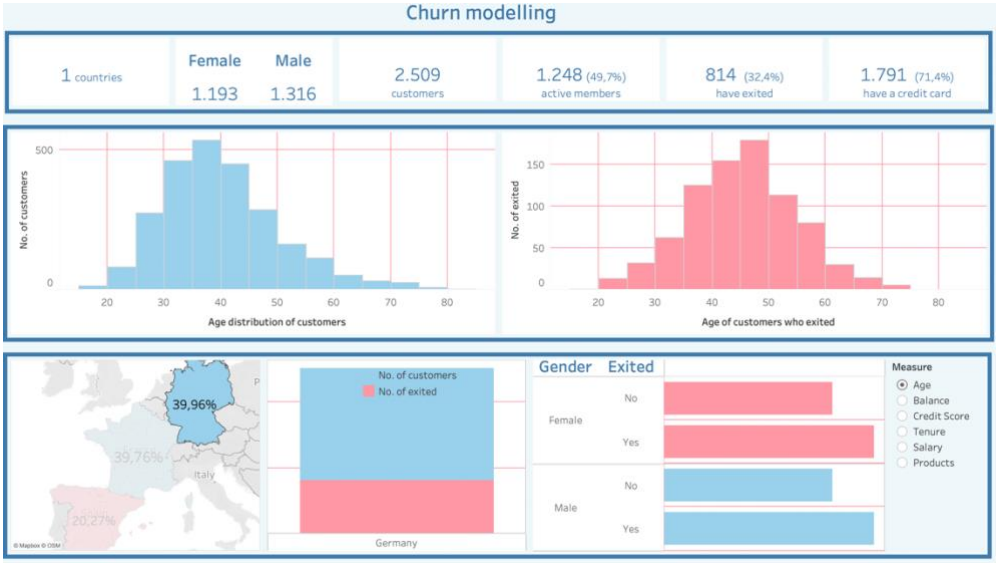
2.5.2 Tableau

The data are published online at *Tableau Public*: https://public.tableau.com/app/profile/sanja.srdanovic/viz/Churnmodellingprojectdashboard/Dashboard1?publish=yes and saved as a *Tableau Workbook*.



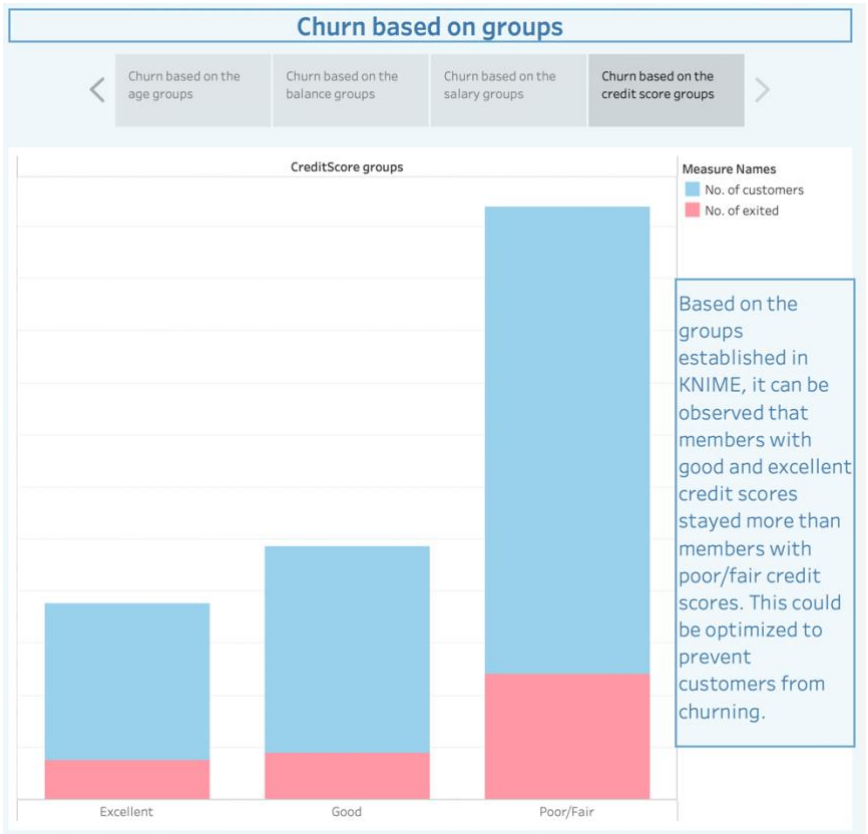
In the dashboard, in the upper part, we can see the numbers and some features of the customers, also the number of active members, members who exited and who possess a credit card is displayed in a percentage as well. Filters are optimized in a manner so that when we click on one bin or one country, the data are altered accordingly. In the second part, the histograms show the number of customers per age distribution, and then the number of customers who churned based on their age. In the third part of the dashboard, there is a map showing the three countries – Spain, France and Germany, as well as a bar chart showing the number of customers and the number of churns in each country. Finally, there is another bar chart showing how many male and female customers left and how many stayed based on a number of measures I set up in the sheet.

The image below shows how filtering works - when we click on Germany for instance, all the other graphs are optimized, and the numbers are changed respectively so that information just for this country are displayed. When we click again, we get the first view with the overall data.



The graphs show that *Age, Balance* and *Number of products* have an influence on churn. When it comes to the *Number of products*, there are fewer people who churned. For *Age* and *Balance*, there are slightly more people who exited than people who stayed. If we have a look at age bins, we can see that people from 40 to 50 leave the most. If we compared the countries, in Germany and France there is 39% of churn more than in Spain which has 20% of churn.

Additionally, I created a *Story* in *Tableau* based on the groups established in *KNIME*. Looking at the *Age groups*, the most churn is in the age range from 40 to 50 years. When it comes to *the Balance groups*, it can be observed that members who exited had a medium balance. Based on the *Salary groups* created in *KNIME* - low, mid and high, it can be observed that members with low and mid salaries churned slightly more than customers who had higher salaries. Finally, it can be observed that members with good and excellent *Credit scores* stayed more than members with poor/fair credit scores. This could be optimized to prevent customers from churning. An example of the story points is shown below, and the whole story is available at the *Tableau Public*:
https://public.tableau.com/app/profile/sanja.srdanovic/viz/StoryChurnbasedongroups/Story1?publish=yes



2.6 Ausblick

The analysis in *KNIME* shows that churn can be very well predicted, and the best accuracy is produced when employing *Random Forest*, it was possible to get predictions for the classification of customers churning or not with quite a good accuracy - 86,4 %. The factors that have the most impact on the churn are: *Gender* - more women than men churned, *Age* – depending on the age group, people who exited the most were between 40 and 50 years old, *Balance* – people with medium balance, *Number of Products* – customers with more products churned more, and for *Credit score* – members with worse credit score exited more than people with good or excellent credit scores.

Based on this analysis, banks could optimize some of their services in order to stop their customers from leaving. What could be also very helpful for the analysis and for minimizing the number of customers who left, is to have more detailed insights into time - when exactly people are churning. Based on these pieces of information, we could also create a forecast to see what the churn would look like in the upcoming period based on the given features.

The procedure described here is a good starting point for practising how to create a workflow in *KNIME* and examine different machine learning models in order to see which of them performs the best. The results predict the churn of customers based on various features quite well, and a comparable procedure could be applied to other datasets with a similar topic.