IBM Introduction to Machine Learning

Specialized Models: Time Series and Survival Analysis

Final Project – Esteban Carboni

Main objectives of the analysis

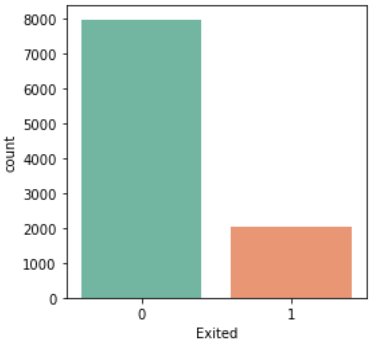
The main objective of this analysis is to identify which variables cause bank customers to churn and how long would they remain a customer. The models for achieving these goals will be focused on Survival Analysis: The Cox Proportional Hazards Model and a Kaplan-Meier Curve approach. Understanding both significances, as well as relative magnitudes of the coefficients will give information to stakeholders in which variable have the largest impact on survival risk.

Data description

The dataset contains 10.000 rows and 14 attributes. First, we drop needless variables like "RowNumber", "CustomerId" or "Surname", keeping significant ones only. Our target variable is "Exited", a binary variable displaying 0 for current customers and 1 if the customer has churned. Here is a summary of the attributes.

	count	mean	std	min	25%	50%	75%	max
CreditScore	10000.0	650.528800	96.653299	350.00	584.00	652.000	718.0000	850.00
Age	10000.0	38.921800	10.487806	18.00	32.00	37.000	44.0000	92.00
Tenure	10000.0	5.012800	2.892174	0.00	3.00	5.000	7.0000	10.00
Balance	10000.0	76485.889288	62397.405202	0.00	0.00	97198.540	127644.2400	250898.09
NumOfProducts	10000.0	1.530200	0.581654	1.00	1.00	1.000	2.0000	4.00
HasCrCard	10000.0	0.705500	0.455840	0.00	0.00	1.000	1.0000	1.00
IsActiveMember	10000.0	0.515100	0.499797	0.00	0.00	1.000	1.0000	1.00
EstimatedSalary	10000.0	100090.239881	57510.492818	11.58	51002.11	100193.915	149388.2475	199992.48
Exited	10000.0	0.203700	0.402769	0.00	0.00	0.000	0.0000	1.00

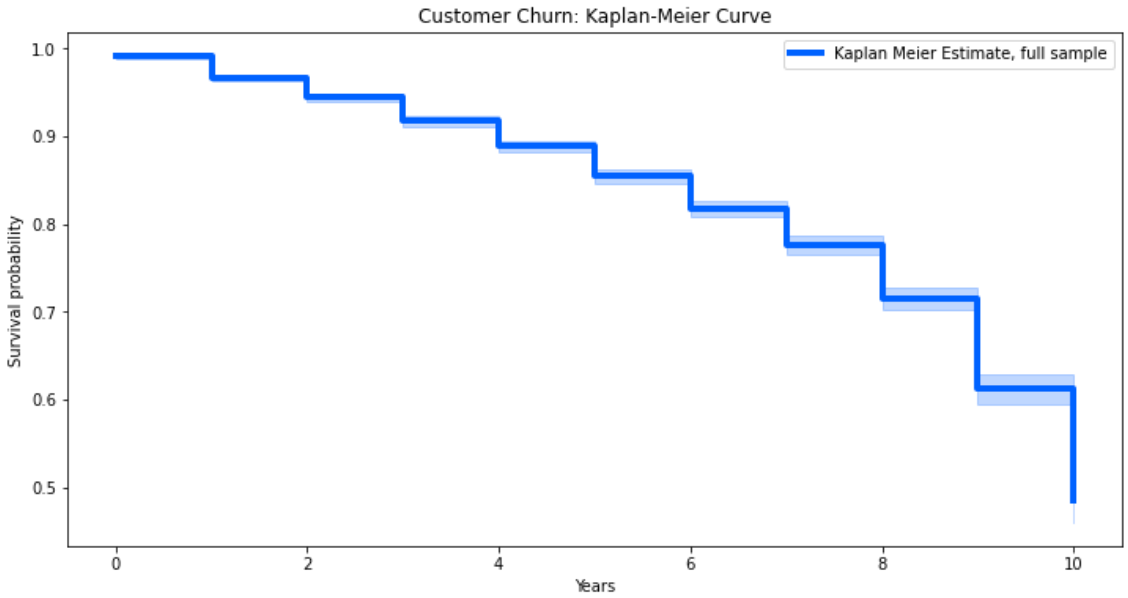
Customer's self-reported annual salary and account balance is shown here. "CreditScore" represents the creditworthiness of an individual. Tenure is an essential variable, shows the years that the customer has stayed loyal to the bank. "NumOfProducts" shows the amount of banking products each customer owns. For binary variables "HasCrCard", "IsActiveMember" and "Exited", 0 represents "NO" and 1 means "YES". Here is a brief sample of the dataset.



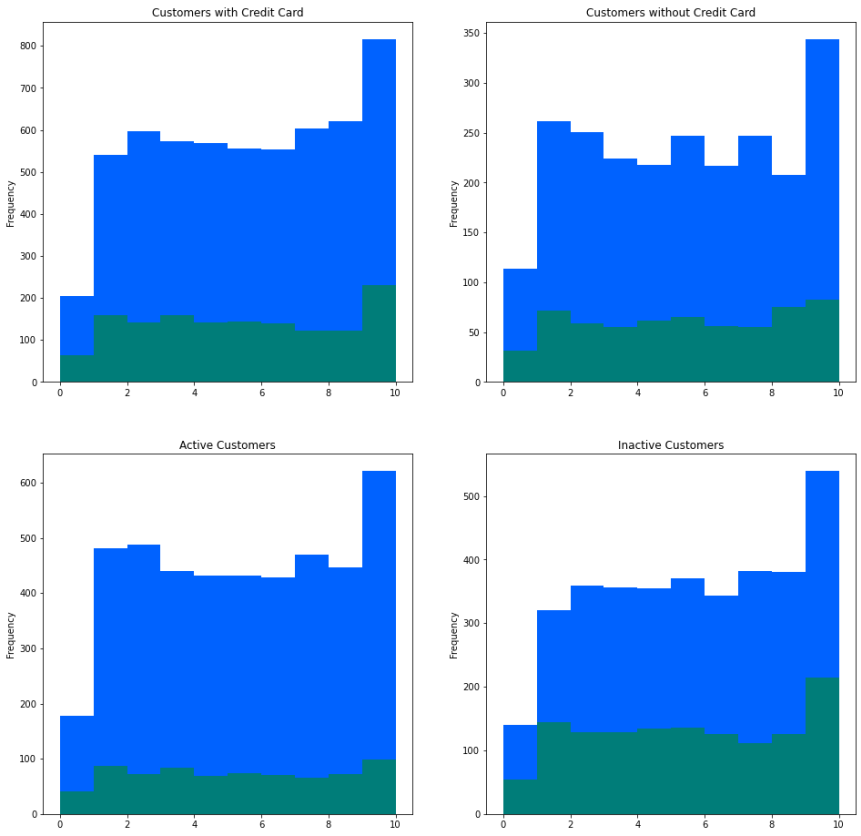
	CreditScore	Geography	Gender	Age	Tenure	Balance	NumOfProducts	HasCrCard	IsActiveMember	EstimatedSalary	Exited
9186	652	Germany	Female	58	3	116353.20	2	0	1	193502.90	0
1481	592	Spain	Male	38	8	0.00	2	1	0	180426.20	0
7218	757	France	Male	36	7	144852.06	1	0	0	130861.95	0
6775	469	France	Female	48	5	0.00	1	1	0	160529.71	1
9803	673	France	Male	31	1	108345.22	1	0	1	38802.03	0
3981	638	France	Male	24	1	0.00	2	0	1	162597.15	0
5773	523	Spain	Female	36	8	113680.54	1	0	0	13197.44	0

Model deployment

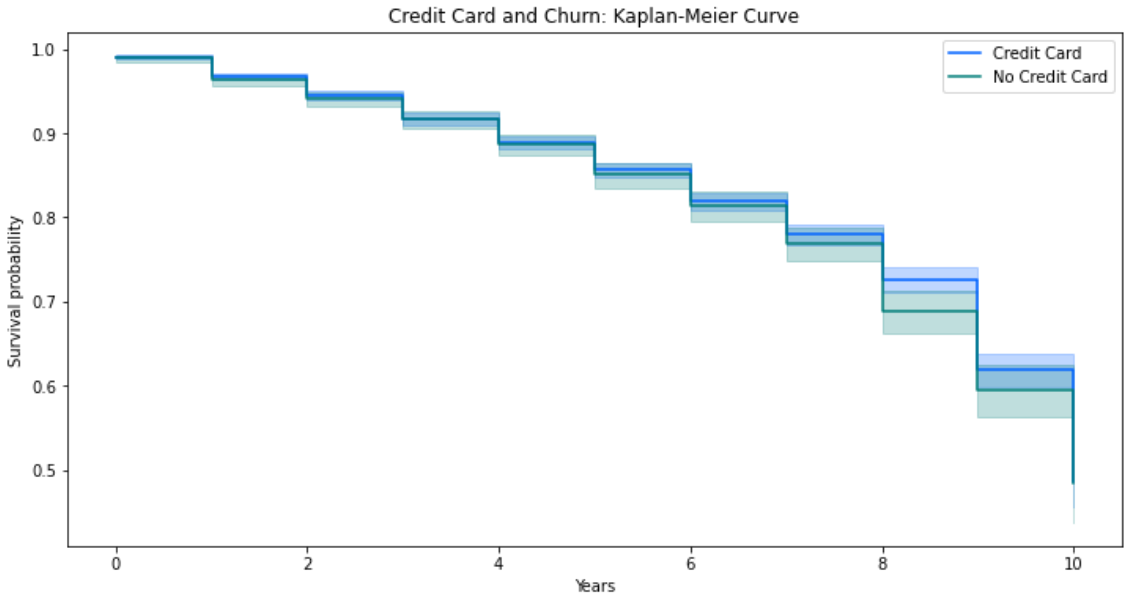
At a first glance, we fit a simple Kaplan-Meier Curve on our data. This represents a simple non-parametric visualization of survival likelihood function. Tenure and Exited variables are integrated into the model. As we could suspect, survival probability decreases over the years.



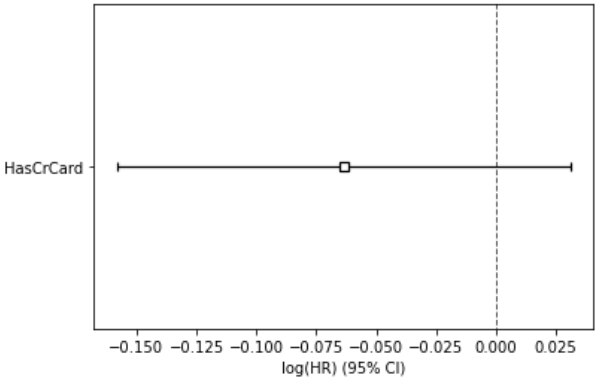
As we continue to examine the survival function, we may want to relate survival risk with features, or characteristics of our customers. We will look at differences in survival risk for customers who have a credit card and customers without a credit card. Also, active customers and inactive customers are analyzed. A simple histogram for each category is plotted, looking at differences between churned and not-churned subsamples.



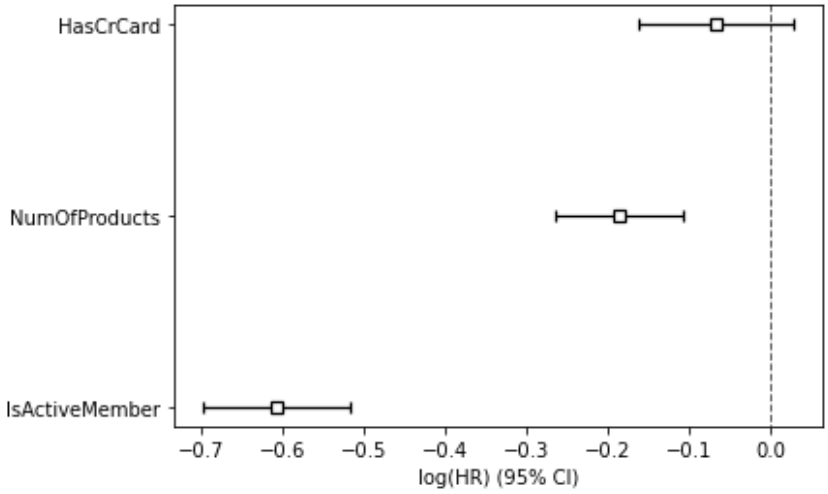
Now we plot another Kaplan-Meier Curve, including the survival probability based on the ownership or not of a credit card.



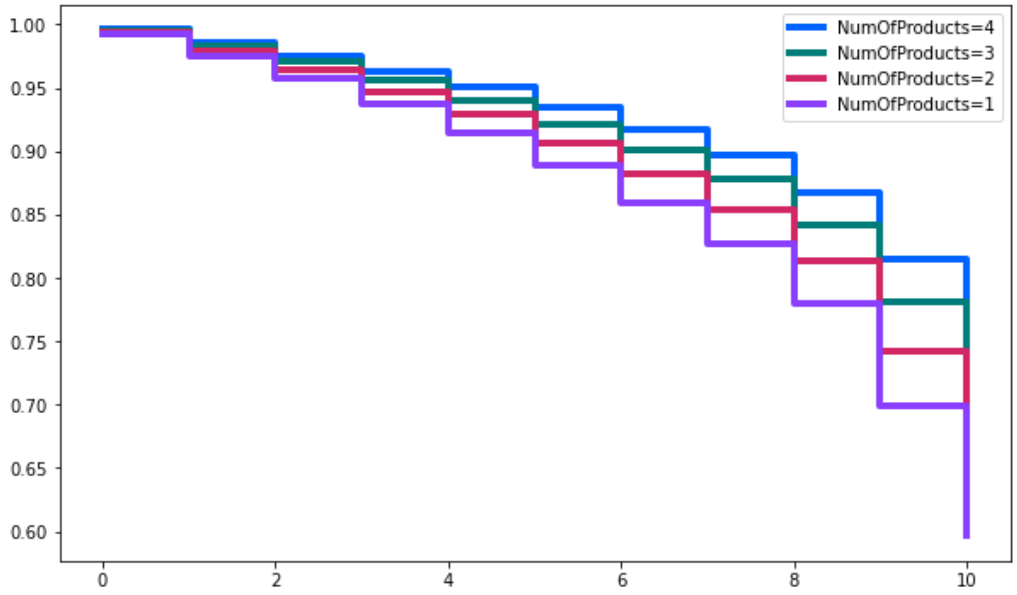
Having looked at our data and related Kaplan-Meier curves, we can formalize the analysis by running survival regression. This Cox model estimates a baseline hazard rate, and assumes features impact this hazard rate proportionally. We fit a Cox proportional hazard model using the "HasCrCard" variable. The Cox model also allows us to plot coefficient estimates to assess their significance.

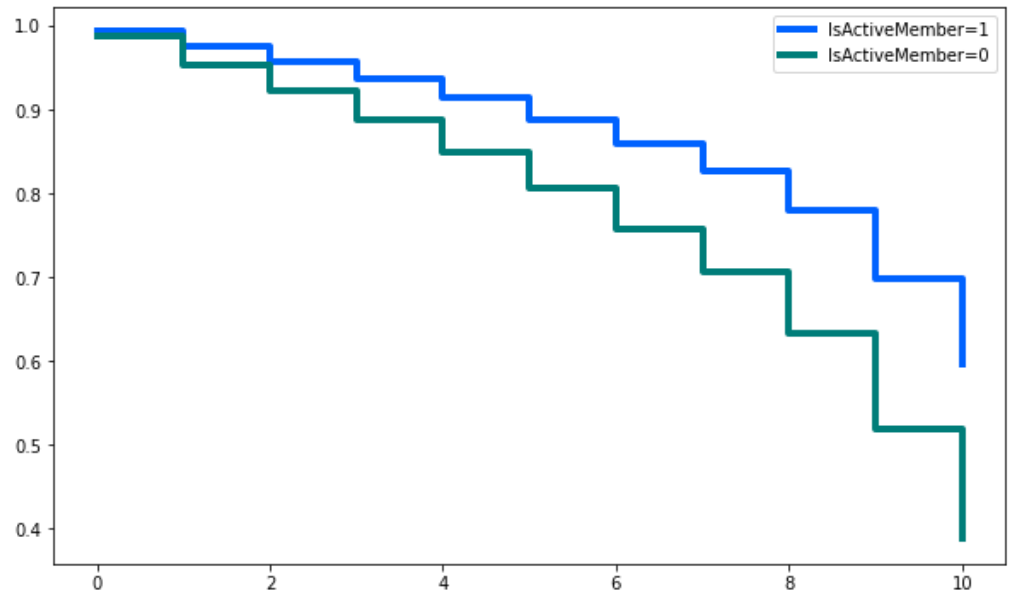


By getting the regression output for the estimated cox model we find out that the result is not significantly different from zero. To compare the impact across variables, we include a few more variables: number of products and whether or the customer is an active member. This allows us to compare different magnitudes and assess which variables have larger influence on survival risk. The IsActiveMember variables seems to have the largest impact on survival risk.



Finally, we plot the partial effects on outcome for number of products and the "IsActiveMember" variables.





Key findings

To avoid customers from churning, bank could offer stimulus or promote actions to turn inactive members to active, because banking inactivity carries the worst survival rate. Then offering more products could lead to less churning. Owning or not a credit card is not so relevant when analyzing banking churn.

Suggestions for next steps

Models should be revisited incorporating other variables such as age, credit score, account balance or estimated salary from the customers to achieve a better model. This report could be inaccurate and present possible flaws because the dataset is split in different years and should be divided in monthly periods for better analysis and model performance. Python notebook code can be found on GitHub: [IBM-Introduction-to-Machine-Learning/Specialized Models Final Project.ipynb at master · estebancarboni/IBM-Introduction-to-Machine-Learning (github.com)]