Predicting Customer Churn in Banking Industry Using Machine Learning

Presenter: Mengdi Hao

Video: https://www.youtube.com/watch?v=Uu5CIWsGsVc

Motivation & Research Question

Motivation

- In recent decades, options for customers to store their money has been rapidly increasing.

→ Customer churn has been one of the top issues for many banks!

Research Questions

- What factors indicate a customer churning or not?
- Can banks detect who are more likely to churn and take measures to those customers?

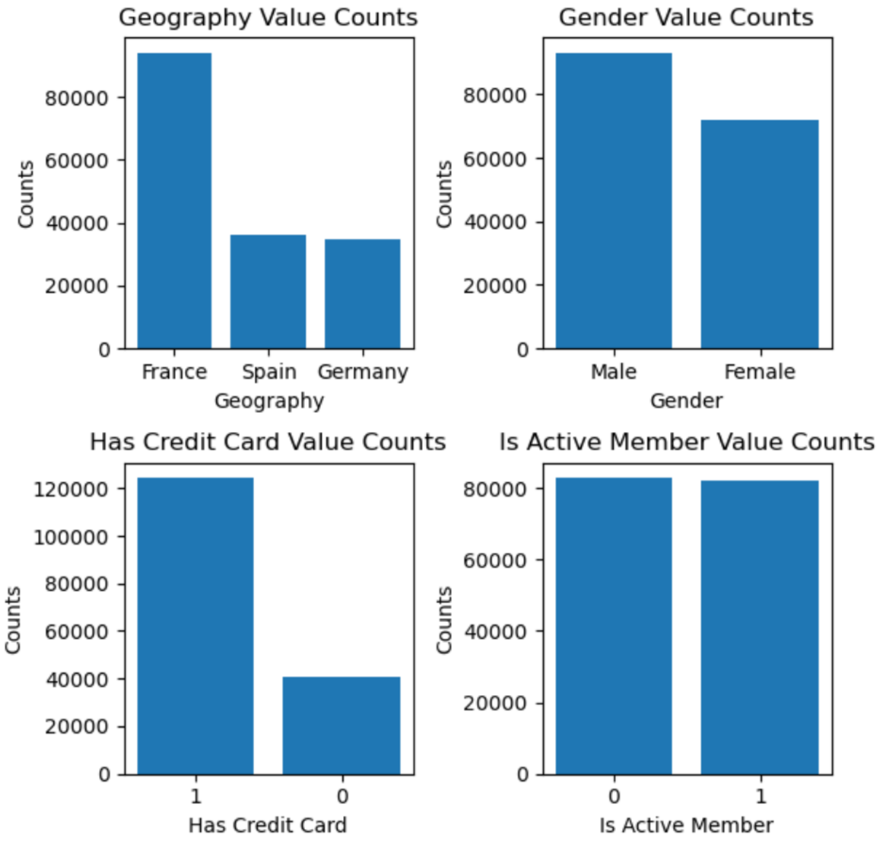
Dataset Introduction

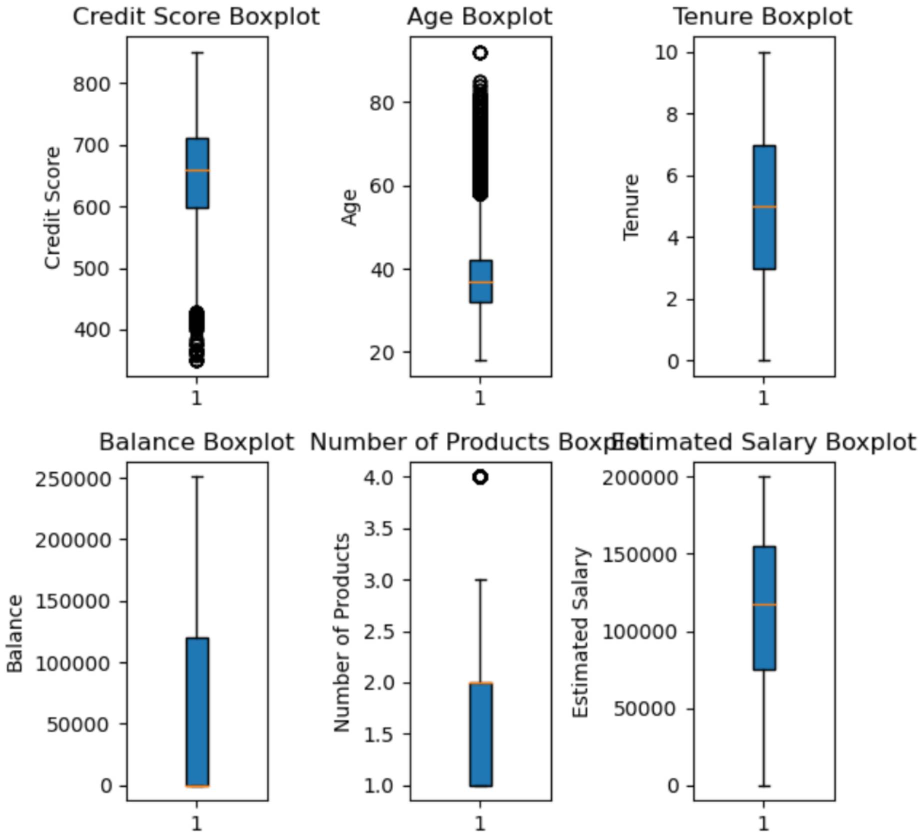
- Generated from a deep learning model: Kaggle link
- Dimension: 165,034 observations, 11 variables
- Variables:

  credit score, geography, gender, age, tenure, balance, number of products, has credit card or not, is active member or not, estimated salary, exited or not

- No missing values

Data Processing – Detect Outliers

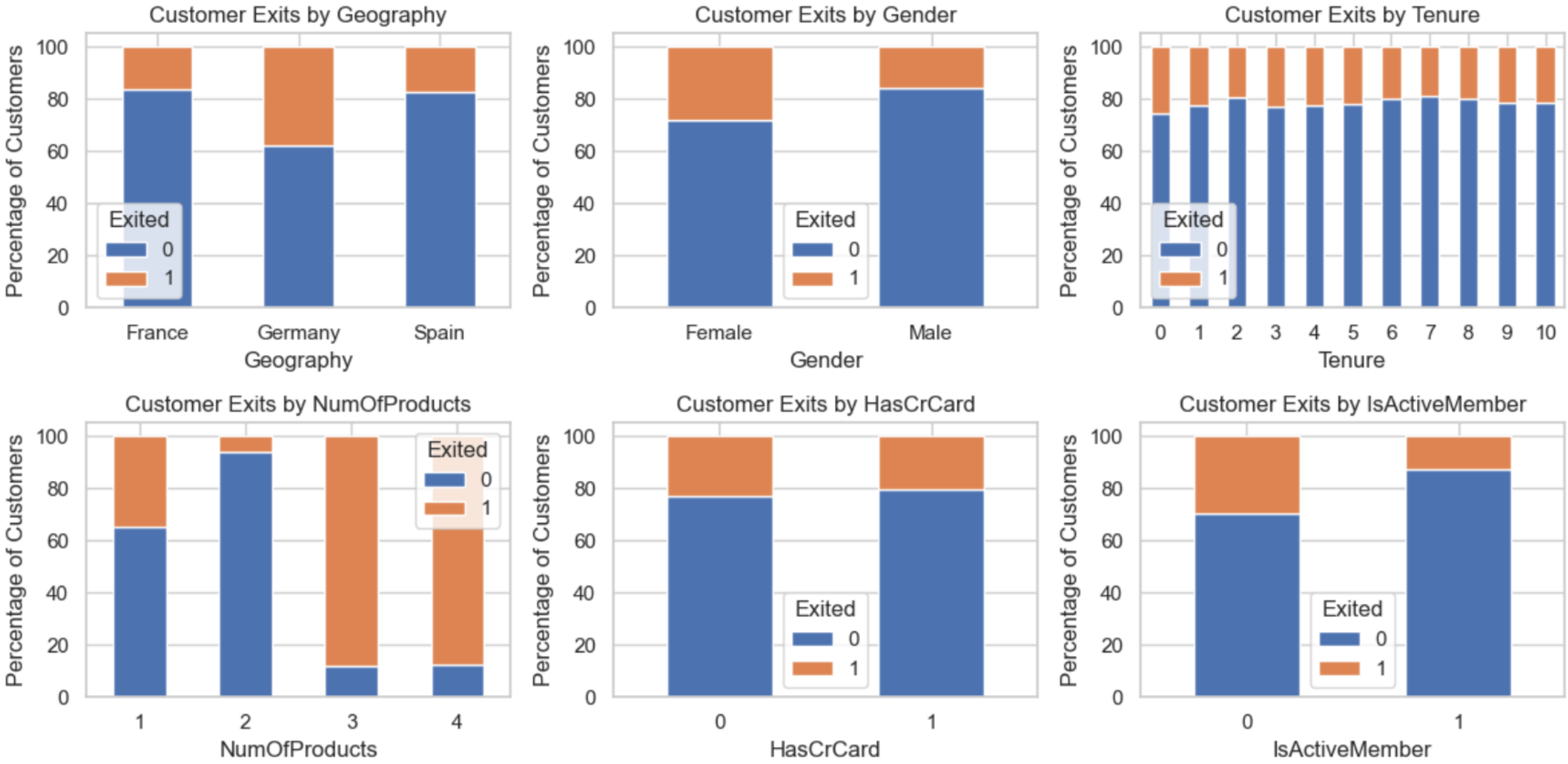




- Categorical

- Numerical

Data Exploration – Visualization



Customer Exits by Geography

Customer Exits by Gender

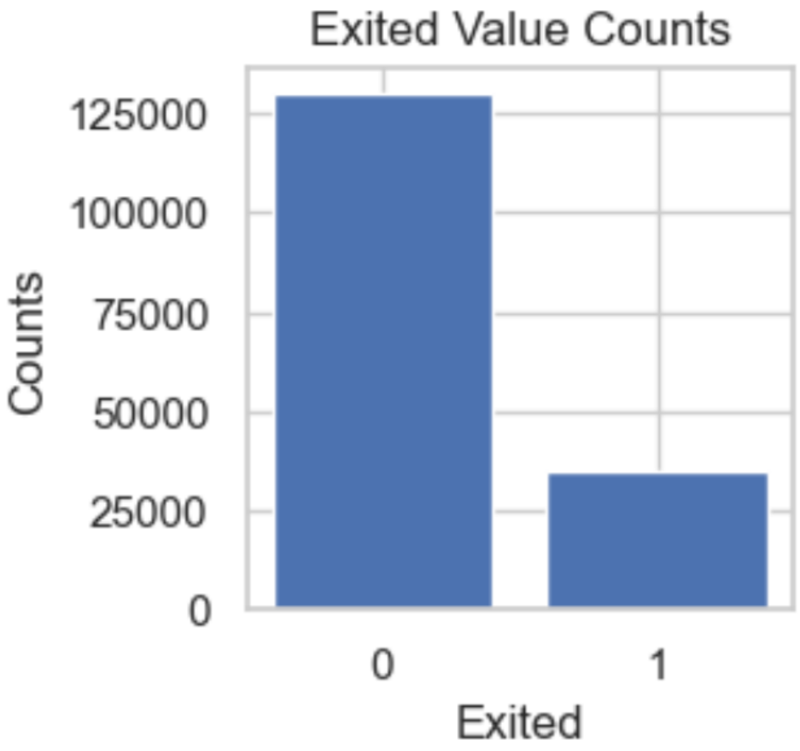
Customer Exits by Tenure

Customer Exits by NumOfProducts

Customer Exits by HasCrCard

Customer Exits by IsActiveMember

Data Exploration – Visualization



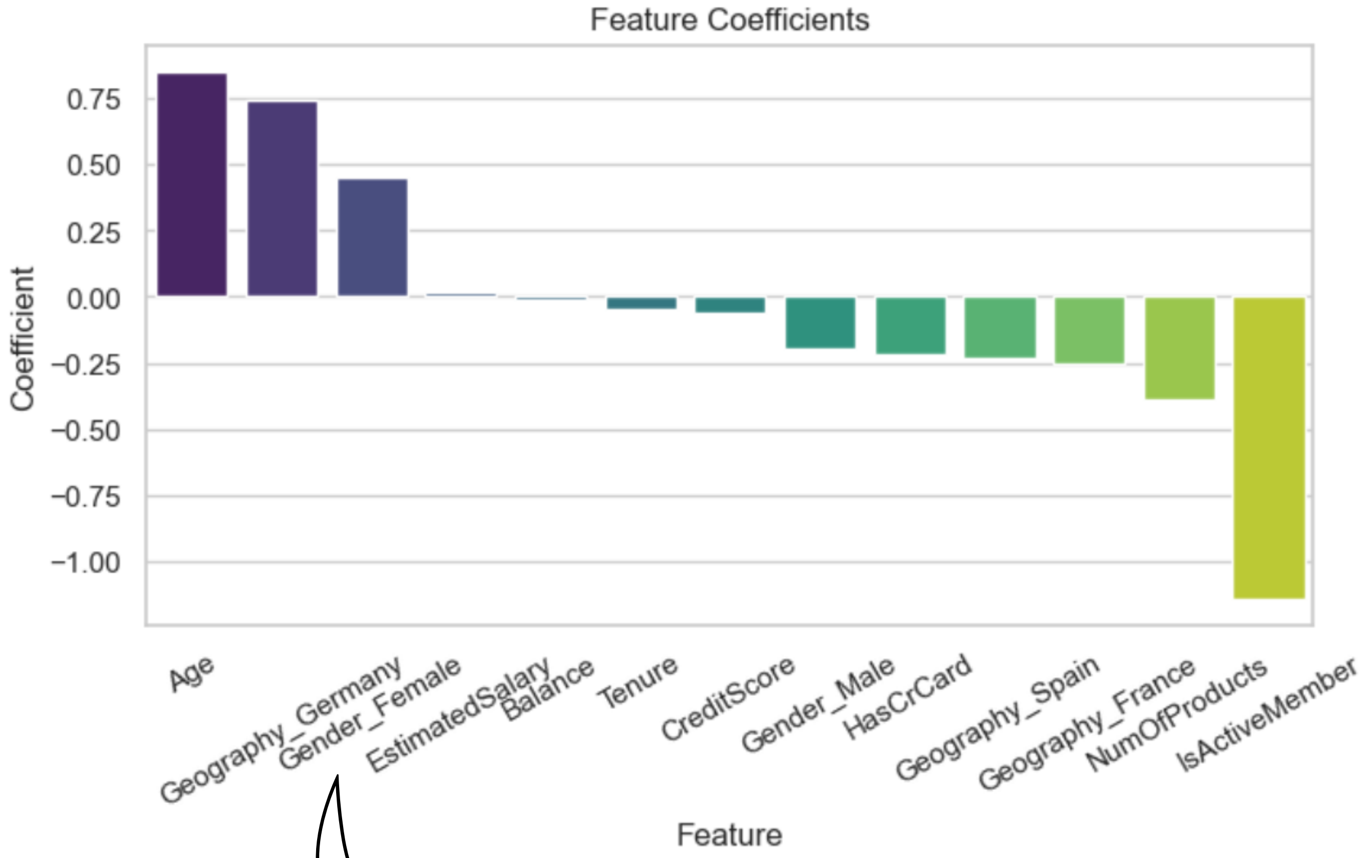
High imbalance! Need to be handled!

- Apply down-sampling technique only on training data;
- After: 49,086 observations

Supervised Learning – Logistic Regression

Model	CV AUC score
Default LR	0.8188
Tuned LR	0.819
Improvement after tuning	*0.02%*

Parameter	Meaning
C	regularization strength
penalty	regularization type
solver	optimization algorithm

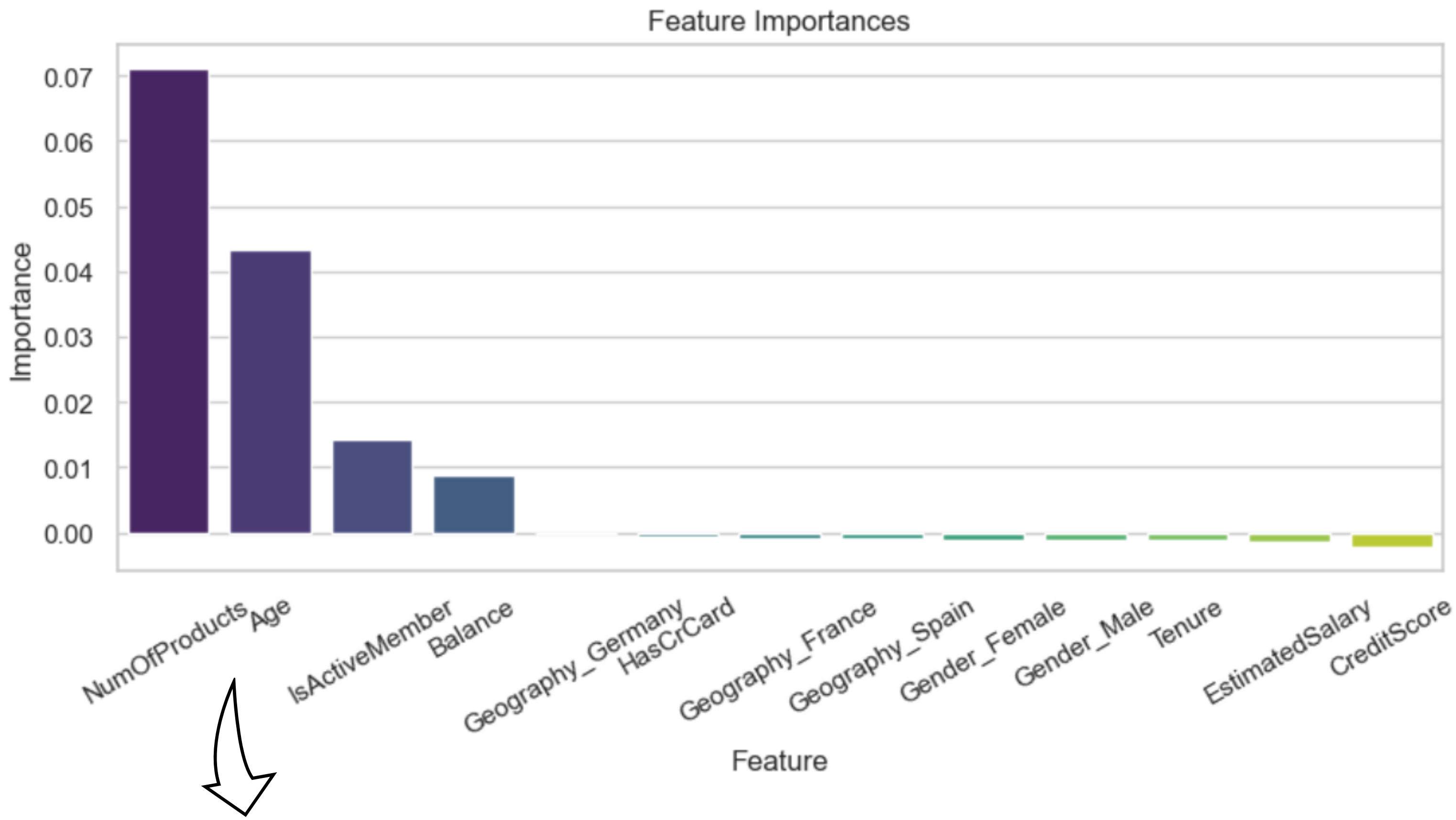


Generally consistent with insights from correlation matrix.

Supervised Learning – KNN

Model	CV AUC score
Default KNN	0.8417
Tuned KNN	0.8691
Improvement after tuning	*2.74%*
Improvement on optimal LR	*5.01%*

Parameter	Meaning
n_neighbors	number of nearest neighbors to consider
weights	way to weight the neighbors' vote
metric	define the distance metric used
p	specify the power parameter of the Minkowski metric

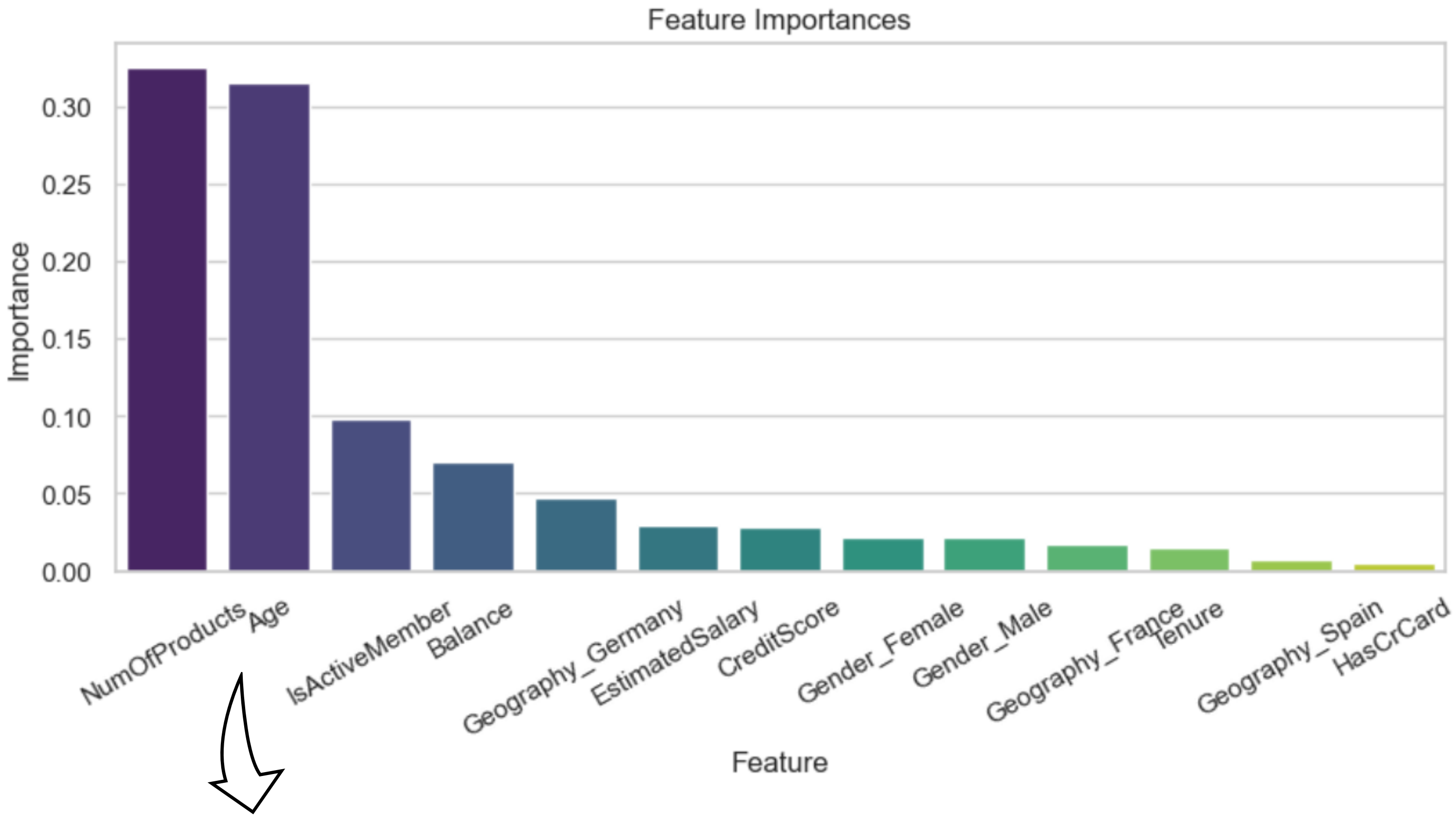


Generally consistent with logistic regression.

Supervised Learning – Random Forest

Model	CV AUC score
Default RF	0.8727
Tuned RF	0.8849
Improvement after tuning	*1.22%*
Improvement on optimal KNN	*1.58%*

Parameter	Meaning
n_estimators	number of trees
max_depth	maximum depth of tree
min_samples_split	minimum samples required for splitting a node
min_samples_leaf	minimum samples required for a leaf
max_features	number of features to consider for a split

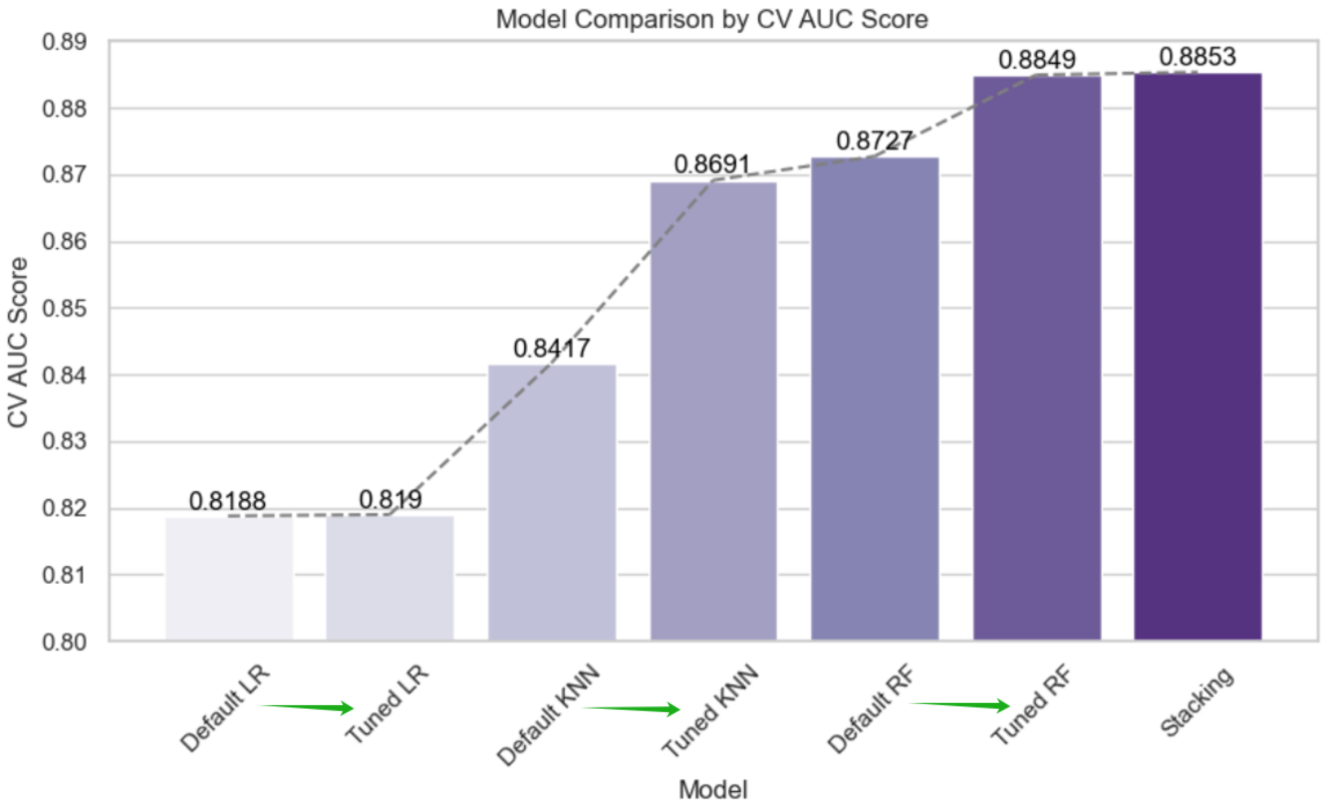


Generally consistent with logistic regression and KNN.

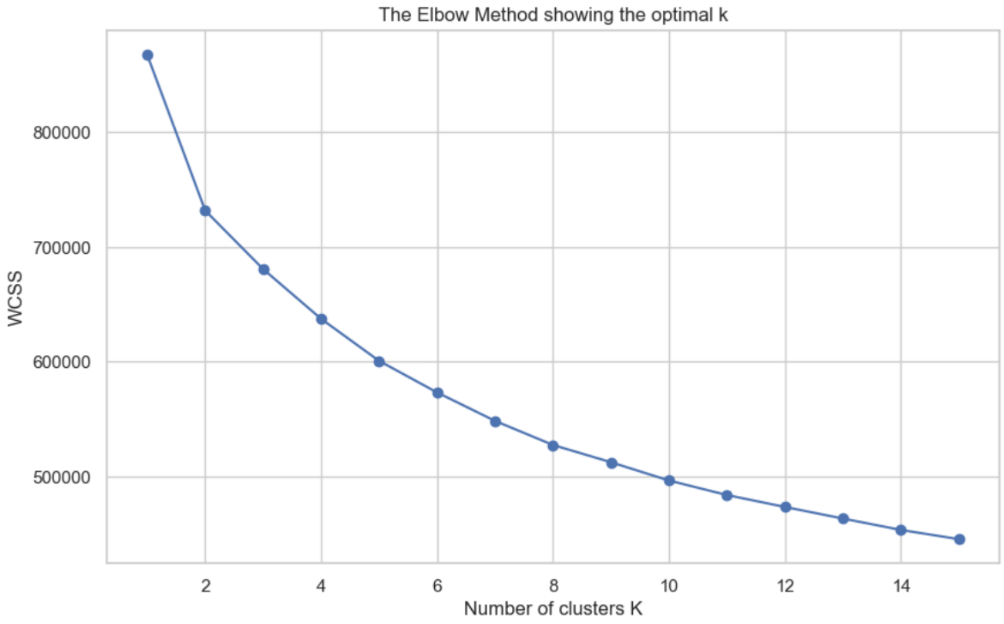
Supervised Learning – Stacking

Model	CV AUC score
Stacking	0.8853
Improvement on optimal RF	*0.04%*

Step by step improvement!



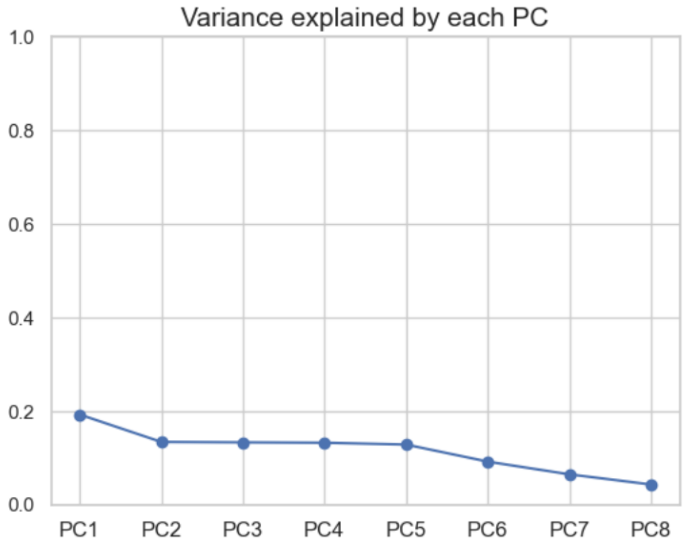
Unsupervised Learning – K means Clustering

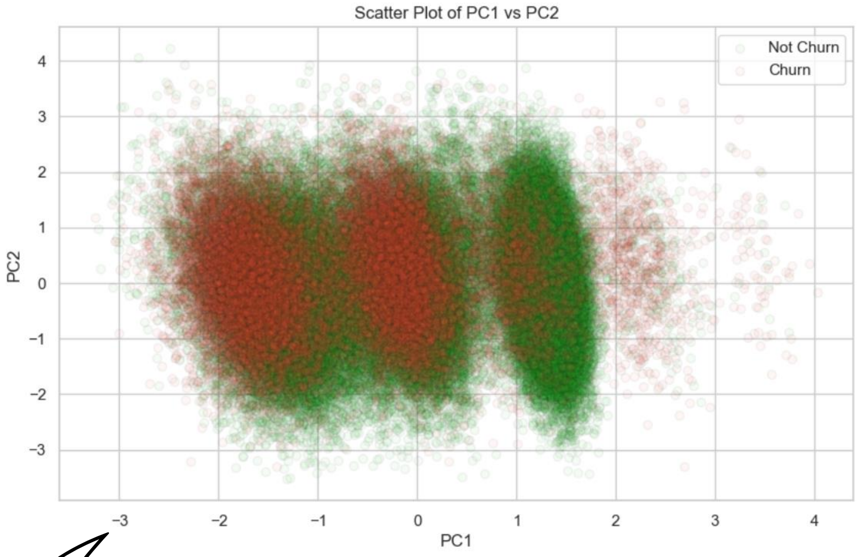


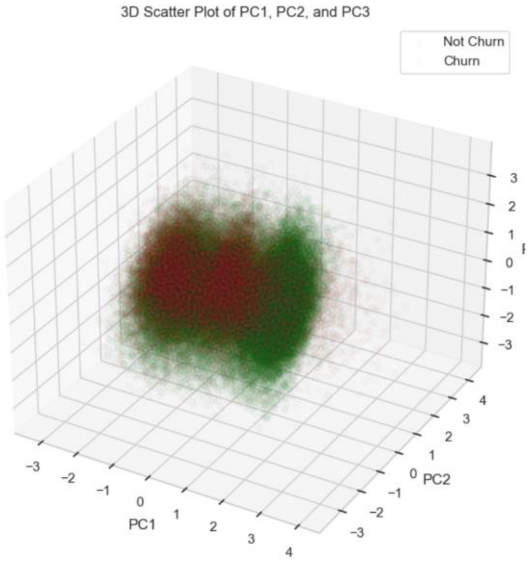
No obvious elbow point!

K	CV AUC score	Improvement?
2	0.8849	×
3	0.8849	×
4	0.8851	×
5	0.8848	×
6	0.8851	×
7	0.8848	×

Unsupervised Learning – PCA



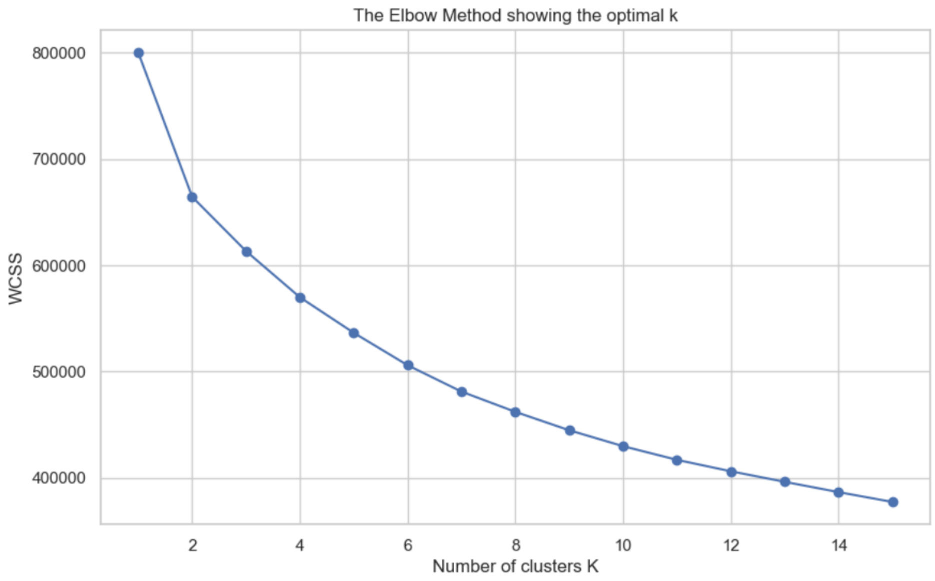




Hard to distinguish the two classes.

AUC score using PCA result: 0.8716, no improvement!

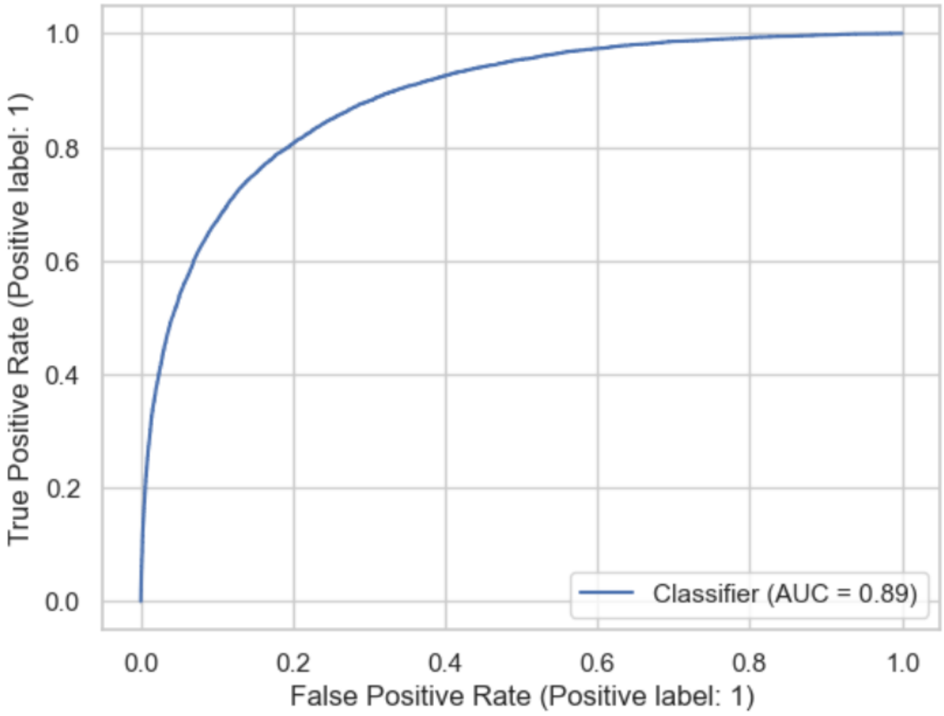
Unsupervised Learning – PCA & Clustering



No obvious elbow point again!

K	CV AUC score	Improvement?
2	0.8727	×
3	0.8729	×
4	0.873	×
5	0.8728	×
6	0.8728	×
7	0.8725	×

Model Evaluation



Classification Report:				
	Precision	**Recall**	**F1-score**	**Support**
0	0.94	0.81	0.87	39133
1	0.52	0.8	0.63	10378
accuracy	0.8			49511
macro avg	0.73	0.8	0.75	49511
weighted avg	0.85	0.8	0.82	49511

Generally, the model performance is good!

Error Analysis – Misclassified Observations

	Age	Balance	NumOfProducts	IsActiveMember	True Label
count	9659	9659	9659	9659	9659
mean	0.315031	0.331976	-0.597615	0.390413	0.215033
std	1.033138	0.991838	0.830157	0.487868	0.410866
min	-2.271851	-0.875434	-1.017052	0	0
25%	-0.352048	-0.875434	-1.017052	0	0
50%	0.2126	0.725191	-1.017052	0	0
75%	0.890178	1.160012	-1.017052	1	0
max	6.08494	3.125753	4.475049	1	1

Seem to be some patterns?

Potential Improvement Strategies:

- Create Interactive Features
- Resampling Techniques
- Hyperparameter Tuning

Conclusion

What factors indicate a customer churning or not?

- *Age;*
- *Geography;*
- *IsActiveMember;*
- *NumOfProducts.*

Can banks detect who are more likely to churn and take measures to those customers?

Yes!

- *ROC-AUC score: 0.8864;*
- *Recall for churn class: 0.80.*