GEOMATIC PREDICT:
ANALYSIS PERFORMANCE REPORT

Part of Data Analytics Platform

powered by GEOMATIC

Churn Prediction

1. What is machine learning

Machine learning at its core is the way to deploy AI. Machine learning is the term used for letting a computer (machine) utilise various statistical techniques to "learn" from historical data (i.e. becoming better at a specific task without being specifically programmed by a human) that in turn creates the best possible predictions on the data for the future. Examples are very wide-spread and can affect all sorts of businesses; for example:

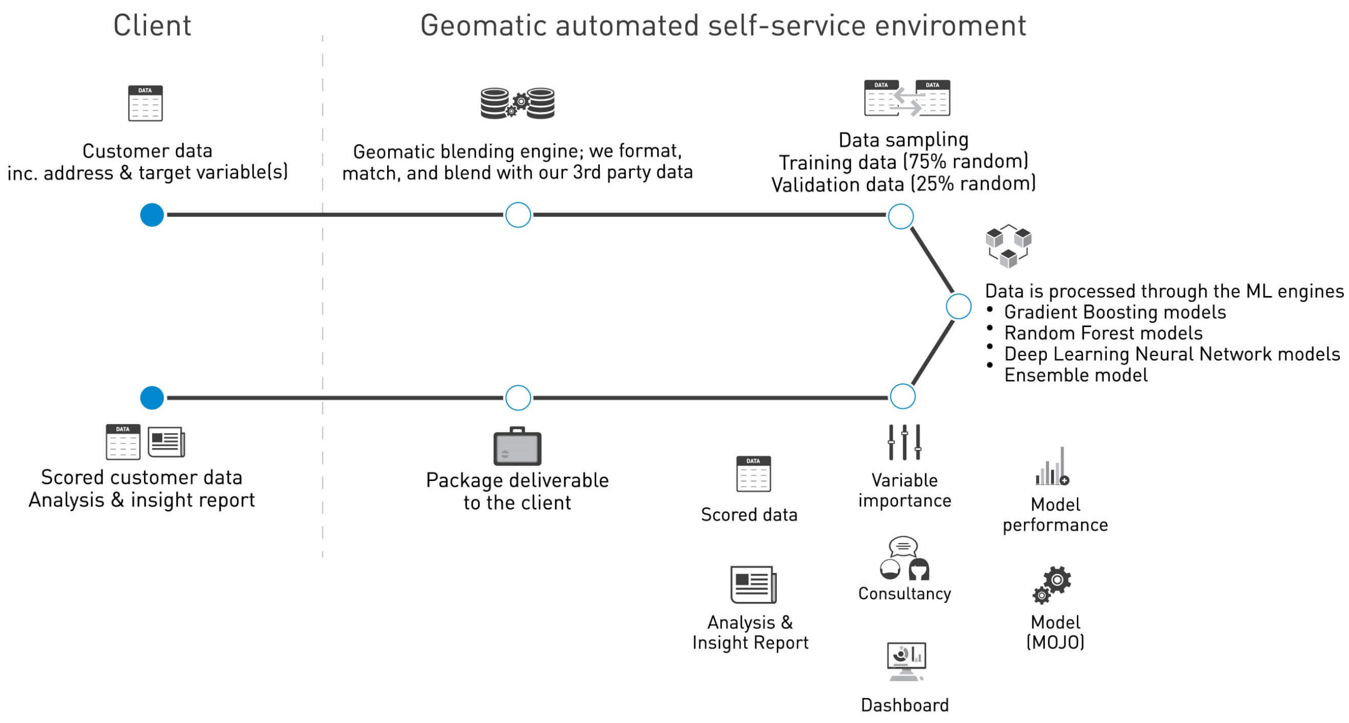
- a retail chain utilises data on shopping-trends to understand how best to plan their inventory
- a hospital applies machine learning on historical patient-flows to understand how best to staff the emergency department/units
- a bank learns from historical fraudulent transactions to identify which current transactions are potentially fraudulent
- a telemarketing division utilises machine learning to better understand who and when to call, in order to optimise sales
- a telco company uses historical churners in a prediction model to help proactively identify future churn and increase loyalty

2. The process behind Geomatic Predict

Using the Geomatic Predict learning engines, we identify the data variables that are the most predictive in relation to the client's need and use case (target variable) e.g. today for churn, tomorrow for risk, etc.

Behind Predict we have implemented a suite of ML tools that utilises the following algorithm techniques:
- gradient boosting models
- random forest
- deep learning neural networks

The process is as follows:
1. the data uploaded by the customer is matched to Geomatic's data via the Geomatic geocoder, then enriched and blended together with all possible data variables
2. the variables uploaded by customers are pre-processed via correlation detection, outlier detection, data imputation, data standardisation and data removal
3. after this the data is randomly assigned into training (75%) and validation (25%) data, where the training data is used for building the models and the validation data is used for validating that the model performs well
4. each algorithm is fed all available variables where they calculate the variable importance
5. using only the variables that stands for 95% of the information value, each algorithm is then used to "re-learn"
6. Since all models are good at different aspects, we combine the optimal combination of models into an ensemble model
7. a score is calculated for all models
8. all models are stored in a MOJO format for future use and operational model deployment
9. the relevant package returns are then created


Client
Geomatic automated self-service enviroment
DATA
Customer data
inc. address & target variable(s)
Geomatic blending engine; we format, match, and blend with our 3rd party data
DATA
DATA
Data sampling
Training data (75% random)
Validation data (25% random)
Data is processed through the ML engines
• Gradient Boosting models
• Random Forest models
• Deep Learning Neural Network models
• Ensemble model
DATA
Scored customer data
Analysis & insight report
Package deliverable
to the client
DATA
Scored data
Variable
importance
Model
performance
Analysis &
Insight Report
Consultancy
Model
(MOJO)
Dashboard

3. Overview summary

Below you can see the match counts for the provided data set.

Match rate	100%
Number of matched customers	61.226
Removed customers due to quality of customer data	11
Total number of customers	61.374

Below you can see the number of records that were used in the models.

Total number of records used in the models	61.215
Number of records in training data	45.912
Number of records in validation data	15.303

3.1 Removed customer data

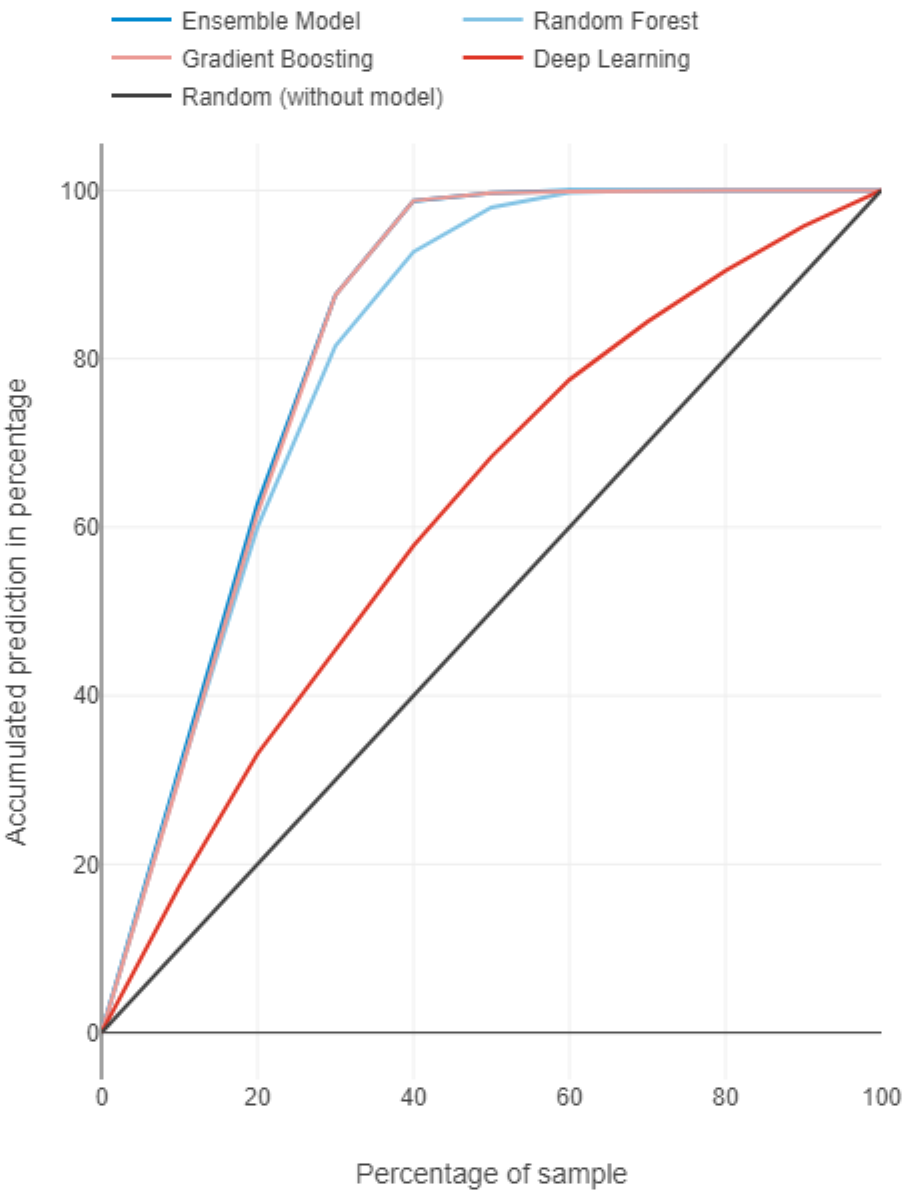
Below you can see the customer variables that have been removed from the model and the reason for the removal.

Variable	Reason for removal
Nps_Feedback_Score (customer variable)	Constant or less than 5% variance

4. Geomatic Predict: Performance overview

Based on measuring the training data to the validation data sets, the table below describes the various performance of the models, showing which model has performed the best with the provided data set. To measure the performance we have used AUC (area under the curve), the closer to 1,0 (100%) the better the prediction.

Model	AUC (in %)
Ensemble Model	90,45%
Gradient Boosting	90,40%
Random Forest	87,94%
Deep Learning	63,65%

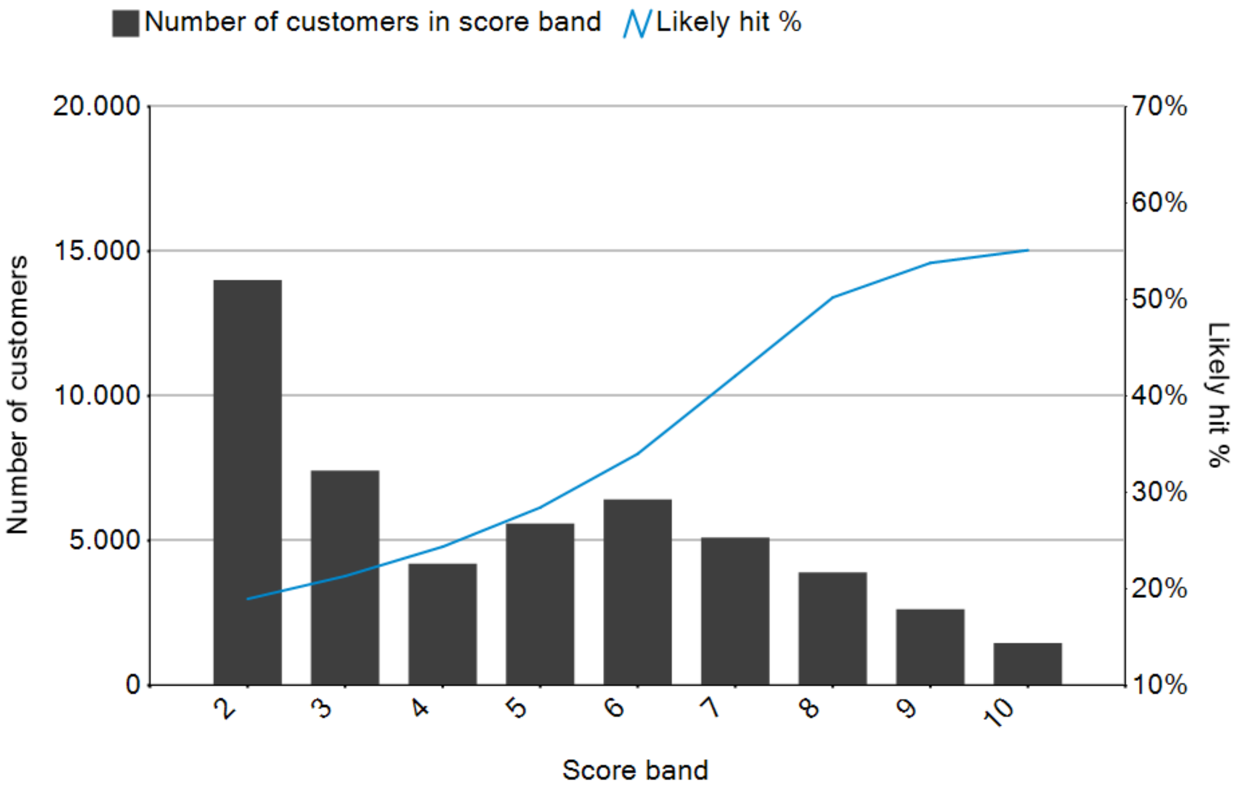


4.1 Assigning the model to full data set

This section describes how the ensemble model performs on the full dataset of customers. In the previous section we computed the model's prediction, and based on this we calculated score bands and hit rates for the validation sample. The hit rate, in the validation sample, is the percentage of customers that was correctly predicted to be "1" by the model.

The next step is to identify the customers in the full dataset (training + validation) that have a predicted likelihood to be equal to the target measure ("1"). Therefore, we select all customers who do not have the input target value equal to "1". Then we apply the model score bands and hit rates to this set of customers. The higher the score band the higher the likelihood that the customer has a value equal to "1", i.e. is a likely hit.

The below graph shows these selected customers from the full dataset and divides them into score bands presented by the height of the bars. The line describes how many percent of the customers in each score band are estimated to be likely hits i.e. their value is close to "1".



5. Best performing variables

5.1 Variable groups

Based on the Geomatic Predict engines, we have identified the data variables and their groups that are the most predictive in relation to your need. The table below shows the results of the ensemble model and importance of each variable group. The most predictive variable groups are on the top and the least predictive variable group on the bottom.

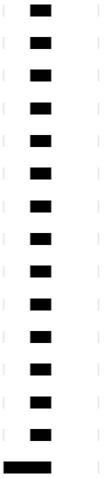
Variable Group	Importance
Geomatic Models: Demographic	41,9%
Geomatic Models: Economic	14,0%
Address Location Admin	13,4%
Customer Variables	8,0%
Geomatic Models: Household	6,5%
Property Data 1-2-1 view	6,1%
Geomatic Models: Attitudes / opinions	2,3%
Geomatic Segmentation conzoom	2,2%
Geomatic Discretion Variables: Property	1,7%
Geomatic Models: Risk (Insurance)	1,6%
Geomatic Models: Property	1,1%
Geomatic Discretion Variables: Economic	0,9%
Geomatic Discretion Variables: Demographic	0,3%

5.2 Variables

The table below shows the results of the ensemble model and importance of each variable used. The most predictive variables are on the top of the table and the least predictive variables on the bottom.

Variable	Importance
Urbanisation model / habitation zone	13,4%
Principal Components Employment	7,9%
Principal Components Life-phase (distribution)	5,2%
Principal Components Highest personal income level	5,1%
Principal Components Household income level	4,7%
Timeonbook_Months (customer variable)	4,6%
Principal Components Highest completed education level	4,2%
Principal Components Origin - 5 groups	3,8%
Principal Components Oldest person in the household	3,1%
Employment level decile	3,0%
Principal Components Social class	2,9%
Margin (customer variable)	2,7%
Principal Components Access to a car	2,6%
Principal Components Family type (marital status)	2,6%
Principal Components Number of children in the household	2,4%
Standardised Property Age (building)	2,4%
conzoom®type (household)	2,0%
Is employed	1,7%
Principal Components Marital status	1,6%
Principal Components Households wealth level (v2)	1,6%
conzoom® flow - flood risk likelihood	1,6%
Wealth decile (v2)	1,2%
Family type decile	1,1%
Construction year (Discretion)	1,0%
There is too little support for refugees score	0,9%
Social class decile	0,9%
Economic trend score	0,8%
Standardised Average time in education	0,8%
Ownership decile	0,8%
Roof material (building)	0,7%
Latest sold price	0,7%
Renovation at the address	0,6%
Ownership (main owner)	0,6%
Standardised Average age (oldest person in the household)	0,6%
Marital status decile	0,5%
Standardised Average household wealth (v2)	0,5%
Children decile	0,5%
Official tax property validation	0,5%
Personal income decile	0,5%
Customerage (customer variable)	0,5%
Age decile	0,4%
Household income decile	0,4%
Education decile	0,4%
Media score	0,4%
Highest personal income (discretion)	0,4%
Property ownership (discretion)	0,3%
Residential property type decile	0,3%
Property size (discretion)	0,3%
Smoker score	0,3%
Standardised Family score	0,3%
Wealth (discretion)	0,3%
Interest for the environment score	0,3%
Household income (discretion)	0,3%

Variable	Importance
Access to car decile	0,3%
Noproducts (customer variable)	0,2%
Age (discretion)	0,2%
Origin decile	0,2%
Transport score	0,2%
Life-phase model	0,2%
Standardised Company car	0,2%
Digitalisation score	0,2%
AVM market price per	0,2%
AVM market price per m2	0,2%
Employment level (discretion)	0,1%
Standardised Average number of children in the household	0,1%
conzoom®group (household)	0,1%
Standardised Energy - estimated natural gas usage	0,1%
Ownership type	0,1%



6. Variable list by model

The table below lists the top 20 Geomatic data variables within each of the three models' engines that the models have identified to have the most predictive value.

Random Forest

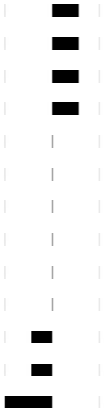
Name	Importance
Urbanisation model / habitation zone	13,8%
Timeonbook_Months (customer variable)	4,4%
Standardised Property Age (building)	3,5%
Principal Components Employment	3,2%
Margin (customer variable)	3,1%
Principal Components Life-phase (distribution)	2,8%
Employment level decile	2,8%
Principal Components Highest completed education level	2,7%
Principal Components Highest personal income level	2,2%
Principal Components Origin - 5 groups	2,2%
conzoom®type (household)	2,1%
Principal Components Household income level	2,1%
conzoom® flow - flood risk likelihood	1,9%
Construction year (Discretion)	1,7%
Marital status decile	1,6%
Wealth decile (v2)	1,6%
Principal Components Social class	1,5%
Children decile	1,5%
Principal Components Households wealth level (v2)	1,5%
Personal income decile	1,4%

Gradient Boosting

Name	Importance
Urbanisation model / habitation zone	26,5%
Principal Components Employment	12,1%
Timeonbook_Months (customer variable)	9,5%
Employment level decile	6,3%
Margin (customer variable)	5,1%
Principal Components Origin - 5 groups	5,1%
conzoom®type (household)	4,0%
Is employed	3,7%
Standardised Property Age (building)	3,7%
conzoom® flow - flood risk likelihood	2,8%
Family type decile	1,9%
Wealth decile (v2)	1,8%
There is too little support for refugees score	1,4%
Economic trend score	1,4%
Principal Components Oldest person in the household	1,4%
Construction year (Discretion)	1,2%
Latest sold price	1,2%
Social class decile	1,2%
Standardised Average time in education	1,1%
Roof material (building)	1,1%

Deep Learning

Name	Importance
Principal Components Highest personal income level	13,2%
Principal Components Life-phase (distribution)	12,7%
Principal Components Household income level	12,0%
Principal Components Highest completed education level	10,1%
Principal Components Employment	8,3%
Principal Components Family type (marital status)	6,7%
Principal Components Oldest person in the household	6,7%
Principal Components Social class	6,6%
Principal Components Access to a car	6,5%
Principal Components Number of children in the household	6,4%
Principal Components Origin - 5 groups	4,1%
Principal Components Marital status	3,5%
Principal Components Households wealth level (v2)	3,3%



7. Customer data pre-processing

The variables uploaded by customers are pre-processed with the following methods before they are used in the ML engines:

1. initial processing of rows and variables
   - row shuffling
   - removal of variables with more than 20% missing values
   - variable classification into categorical and continuous variables
   - removal of sequential variables
   - removal of constant variables and variables with near zero variance
   - removal of duplicate variables
   - removal of rows with more than 40% missing values

2. removal of highly correlated variables

3. data imputation on missing data

4. outlier detection and removal

5. standardisation of numeric variables

8. Terminology

SCORE - Score is a value that is being applied to the entire dataset where we rank the data from the highest probability to the lowest probability to be 1.

SCORE BAND - A score band has a range from 1-10 and is calculated by splitting the test-data into decile where the ten percent of the highest scores will form decile number 10 and so on. This is then converted back to the training set, meaning that the spreads could differ.

PRINCIPAL COMPONENTS - When working with data you will most likely find correlations in data, correlations that could cause issues in model-development. Correlations occur in those cases where you might have two income variables, one in the form of deciles and one in the form of average values. When beginning to train a machine learning algorithm it is crucial to remove the correlations beforehand. This is done by a method called Principal Component Analysis (PCA) where we, automatically, calculate a mathematical "principal component" for income, where the possible correlations are discarded, and therefore will not have an effect in rest of the model development process.

VARIABLE GROUPS - The data variables used in the Geomatic Predict can be divided into the following main variable groups.

Address Location - Includes variables that are based on the geographical area where the person lives

Company Data - Includes variables that are related to the business operations at the address, e.g. is there a company registered on the address, etc.

Geomatic Discretion Variables - Includes discretioned variables related to demographics, economics and property

Geomatic Models - Includes various types of standardised, modelled and countinuous variables related to demographics, economics, property, attitudes and opinions

Geomatic Segmentation - Includes Geomatic's consumer segmentation conzoom®

Permissions - Includes advertising protection list Robinson -list

Property Data - Includes information on the property, sourced from OIS

CUSTOMER DATA - Data variables uploaded by the customer

APPENDIX

Below is a list of the variable codes included in your output file.

Variable	Code
Urbanisation model / habitation zone	HOU_HABITATION_CODE
Principal Components Employment	Empl_PC1, Empl_PC2
Principal Components Life-phase (distribution)	LifePh_PC1, LifePh_PC2, LifePh_PC3, LifePh_PC4
Principal Components Highest personal income level	IncHi_PC1, IncHi_PC2, IncHi_PC3, IncHi_PC4
Principal Components Household income level	Inc_PC1, Inc_PC2, Inc_PC3, Inc_PC4
Principal Components Highest completed education level	Edu_PC1, Edu_PC2, Edu_PC3
Principal Components Origin - 5 groups	Origin_PC1
Principal Components Oldest person in the household	Age_PC1, Age_PC2
Employment level decile	PER_EMPL_LEVEL_FAC_CODE
Principal Components Social class	SocGrp_PC1, SocGrp_PC2
Principal Components Access to a car	Cars_PC1, Cars_PC2
Principal Components Family type (marital status)	HouStruc_PC1, HouStruc_PC2
Principal Components Number of children in the household	Kids_PC1, Kids_PC2
Standardised Property Age (building)	UNADR_PRIMUNIT_BLD_AGE
conzoom®type (household)	HOU_CNZTYP_G5_CODE
Is employed	PER_EMPL_ISEMPLD_FRA
Principal Components Marital status	Marsta_PC1, Marsta_PC2
Principal Components Households wealth level (v2)	Wealth_PC1, Wealth_PC2
conzoom® flow - flood risk likelihood	UNADR_CNZ_FLOW_CODE
Wealth decile (v2)	HOU_WEALTH_V2_FAC_CODE