Introduction
- Profit Driven Business Analytics
- Analytical Process Model
- Analytical Model Evaluation

Analytical Techniques
- Data Preprocessing
- Types of analytics
  - Predictive Analytics
    - Linear Regression
    - Logistic Regression
    - Intermezzo: Variable Selection for Linear/Logistic Regression
    - Decision Trees
    - Regression Trees
    - Neural Networks
  - Ensemble Methods
    - Bagging
    - Boosting
    - Random Forests
  - Evaluating Predictive Models
    - Evaluating Ensembles
    - Splitting up the dataset
    - Performance Measures for Classification Models
    - Performance Measures for Regression Models
  - Descriptive Analytics
    - Association Rules
    - Clustering

Business Applications

Uplift Modeling
- Response Modeling
- Effects of a Treatment
- Data Pipeline

Introduction

Difference between *data* and *information* is that data fundamentally is comprised of zeroes and ones, and information implies in addition a certain utility or value to the end user or recipient.

Business Intelligence provides insight by customized reporting. It is an umbrella term that includes the applications, infrastructure, tools and best practices that enable access to end analysis of information to improve and optimize decisions and performance.

Analytics is a catch-all term covering a wide variety of data processing techniques. It is a toolbox containing a variety of instruments and methodologies allowing users to analyze data for a diverse range of well-specified purposes.

1. *Predictive Analytics:* Based on observed variables, the aim is to accurately estimate or predict an unobserved value.
2. *Descriptive Analytics:* Aim at identifying specific types of patterns.
   - Clustering aims at grouping **entities** of similar nature.
   - Association Analysis aims at finding groups of **events** that frequently co-occur.

Predictive	Descriptive
Classification	Clustering
Regression	Association Analysis
Survival Analysis	Sequence Analysis
Forecasting	Text-Mining

Analytics apply to **structured data**.
Rows are typically called observations, instances, records.
Columns are typically called variables, predictors, characteristics, attributes and features.

Specialized techniques exist to deal with unstructured or semi-structured data such as texts, graphs, etc. Given that roughly 90% of all data is unstructured, there is clearly a large potential for these types of analytics to be applied in businesses.

Profit Driven Business Analytics

Analytics is to be adopted in business for *better decision making*, striving for the optimal in terms of maximizing net profit/value resulting from decisions made on insights obtained from the data by applying analytics. It facilitates optimization of the fine granular decision-making activities leading to lower costs or losses and higher revenues and profit.

The quality of data-driven decision making depends on the extend to which the actual use of the predictions, estimates or patterns is accounted for in the development and application of the analytical approaches. We argue that the *actual goal*, that is to generate profits, should be central when applying analytics.

There is a tangible difference between a statistical approach to analytics and a profit driven approach. F.e. actively retaining customers has been shown to be much cheaper than acquiring new ones to replace those who defect. Not every customer generates the same amount of revenues (think: CLV) to a company. It is therefore much more important to detect churn of customers with a large CLV. From a *statistical* perspective no differentiation is made between both high-and low value customers. From an *analytical* perspective the aim would be to steer or tune the predictive model so it accounts for value.

An additional difference concerns the choice between *explaining* and *predicting*. The aim of estimating a model may be either of these two goals:

1. To establish the relation or detect dependencies between different predictors and a target variable.
2. To estimate or predict a target variable as a function of different predictors.

In applications where the aim is to predict, we are essentially not interested in what drivers *explain* how to realize a target variable of certain value (although this may be a useful side result). We mainly wish to predict as accurately as possible. *This is in many business settings the case.*

Predictive Model	
Classification	A classification model partitions observations in sets based on the target variable.
Regression	A regression estimates a continuous target variable.
Survival Analysis	Survival Analysis, in comparison with classification, is mainly concerned with *when* the event will occur rather than *whether* it will occur.
Forecasting	Forecasting or time series modeling techniques allow an accurate prediction of the short-term evolution of demand based on historical demand patterns.

Descriptive Model	
Clustering	Clustering facilitates automated decision making by comparing a new transaction to clusters or groups of historical transactions
Association Analysis	Often applied for detecting patterns within transactional data.

Analytical Process Model

1. Identifying Business Problem
2. Identifying Data Sources
3. Data Selection
4. Data Preprocessing (Cleaning, Transformation)
5. Optional: Feature Generation
6. Optional: Hyper-parameter Optimization
7. Analyze the data with models
8. Interpret, Evaluate & Tune
9. Deploy the model

The objective of applying analytics needs to be unambiguously defined. Defining the perimeter of the analytical modeling exercise requires a close collaboration between data scientists and business experts. Next all source data that could be of potential interest need to be identified. The golden rule is: *the more data, the better!*

Basic exploratory analysis can than be considered using f.e. OLAP facilities for multidimensional analysis, followed by a data-cleaning step to get rid of all consistencies. Additional transformations may also be considered such as binning, alphanumeric to numeric coding, geographical aggregation etc. , as well as deriving additional characteristics that are typically called features.

In the analytics step, an analytical model will be estimated/trained. Once the results are obtained, they will be interpreted and evaluated by Business Experts.

The key is to find unknown yet interesting and actionable patterns that can provide new insights into your data that can then be translated into new profit opportunities.

Once the model has been validated and approved, it can be put into production as an analytics application (= Decision Support System, Scoring Engine). **The process model is iteratively in nature** in the sense that one may have to return to previous steps during the exercise.

The most time-consuming step typically is the data selection and preprocessing step, which usually takes around 80% of the total efforts needed to build an analytical model.

Analytical Model Evaluation

Before adopting an analytical model and making operational decisions, the model needs to be thoroughly evaluated. Depending on the exact type of output, the setting or business environment and the particular usage characteristics, different aspects may need to be assessed during evaluation in order to ensure the model is *acceptable* for implementation.

A number of key characteristics of *successful* business analytical models are defined and explained in the following table:

Characteristics	
Accuracy:	Refers to the predictive power or the correctness of the model. Several evaluation criteria such as hit rate, lift, AUC may be applied to assess the model. It may also refer to the statistical significance, the underlying data needs to be robust and not a consequence of coincidence. We need to make sure the model *generalizes* well and is not *overfitted* to the historical dataset.
Interpretability:	This aspect involves a certain degree of subjectivism, since interpretability may depend on the user's knowledge and skills. The interpretability depends highly on the models' format. Models that allow the user insights in how it obtained certain results are called *white box* models. F.e. decision trees, linear regression, etc. Other models such as random forests and neural nets are called *black box* models.

Characteristics	
Operational Efficiency:	Refers to the time it takes to make a business decision based on the model's outcome. Crucial for certain business applications such as fraud detection and other banking systems. OE also entails the efforts needed to construct the complete analytical process model.
Regulatory Compliance:	A model should be in line and comply regulatory standards.
Economical Cost:	Developing, implementing, deploying and maintaining the model involves significant costs to an organization. External data may be purchased, cloud computing resources may incur large costs etc.

Analytical Techniques

Data Preprocessing

> Data is the key ingredient for any analytical exercise.

Worth mentioning is the *garbage in, garbage out* principle that essentially states messy data will yield messy analytical models. It is therefore of utmost importance that every data preprocessing step is carefully justified, carried out, validated and documented before proceeding with further analysis.

Aggregating the data: The application of analytics typically requires or presumes the data to be presented in a single table. Several normalized source tables have to be joined/merged in order to construct the aggregated, denormalized table. The individual entity can be recognized and selected in the different tables by making use of (primary) keys.

Sampling: Aim is to take a subset of historical data in order to build a model. Key requirements for a good sample are relevance, representativeness and actuality. Choosing the optimal time window of the sample involves a *trade-off between lots of data and recent data*.

Exploratory Analysis: Some initial insights through data visualization. Summarize the data by using descriptive statistics.

Missing Values: Replacement of missing values through *imputation* techniques such as mean/median. Or deletion of the complete row => loss of data.

Outlier Detection and Handling:

- Valid? => capping technique
- Invalid? => treat as missing value

Principal Component Analysis:

A popular technique for reducing dimensionality, studying linear relationships and visualising complex datasets is PCA. It is has its roots in the linear algebra and is based on the concept of constructing an orthogonal basis of the original dataset.

Four properties need to hold:

1. Each principal component should capture as much variance as possible.
2. Variance should decrease each step

3. Transformation should respect the distances between observations and the angle that they form
4. Coordinates should not be correlated with each other

PCA is one of the most important techniques for data analysis. It helps in getting a quick overview of the data and the most important variables in the dataset.

Application areas:

- Dimensionality Reduction
- Input Selection: calculating factor loadings
- Visualisation
- Text mining and information retrieval

The most efficient method to calculate the PCA's is Singular Value Decomposition.

PCA only holds for linear relationships, for non-linear relationships Kernel PCA can be used.

Types of analytics

Predictive Analytics

> In predictive analytics, the aim is to build an analytical model predicting a target measure of interest. The target is then typically used to tune the learning process during an optimization procedure.

Linear Regression

The most commonly used technique to model a continuous target variable.

$$y = \beta_0 + \beta_1 x_1 + \ldots + \beta_k x_k$$

With

$$x_i = variables$$
$$\beta_i = parameters$$

minimized by their sum of squares (often referred as Ordinary Least Squares = OLS). This technique is easy to understand and contributes to operational efficiency. Other, more sophisticated techniques are ARIMA, VAR, GARCH and MARS.

Logistic Regression

When modeling the binary response target two problems arise:

1. Errors are not normally distributed but follow a Bernoulli distribution.
2. No guarantee that the target is between 0 and 1.

Therefore a *bounding function* was introduced:

$$f(z) = \frac{1}{1 + e^{-z}}$$

so that

$$P(y = 1 | x_1, \ldots, x_k) = \frac{1}{1 + e^{\beta_0 + \beta_1 x_1 + \ldots + \beta_k x_k}}$$

Interpretation: The amount of change required for doubling the primary outcome odds equals the so-called doubling amount:

$$log(2)/\beta_i$$

To simplify the optimization, the logarithmic transformation of the likelihood function is taken and the corresponding logit can then be optimized using iteratively re-weighted least squares method.

The logistic regression estimates a linear decision boundary to separate both classes.

Intermezzo: Variable Selection for Linear/Logistic Regression

Criterions for variable selection:

- More interpretable model
- More robust/stable model
- Reduced collinearity
- Operational efficiency
- Economic cost
- Regulations

Model	Test Statistic
Linear Regression:	Student *t*-distribution with n-2 degrees of freedom.
Logistic Regression:	Chi-squared distribution with 1 degree of freedom

For small numbers of variables, an exhaustive search can be used. Otherwise forward, backward or stepwise regression is preferred.

Decision Trees

Decision trees are recursive partitioning algorithms. The calculations can be perfectly parallelized. The top root specifies the *testing condition*. The tree leaves assign the classifications. Implementations differ in the ways they answer the key decisions to build a tree , which are:

- Splitting decision
- Stopping decision
- Assignment decision

Assignment Decision: typically looks at the majority class.

Splitting Decision: Based on the concept of impurity or chaos.

Minimal impurity (= maximal purity) occurs when the variables are all of the same class. Maximal impurity is obtained when both classes represent 50% of the variables. *Decision trees aim to minimize impurity in the data*.

Quantitative Measures of Impurity:

$$C4.5 => entropy: E(S) = -p_G log_2(p_G) - p_B log_2(p_B)$$
$$CART => Gini(S) = -2 p_G p_B$$
$$CHAID => ChiSquared$$

The weighted decrease in entropy thanks to the split is called the *information gain*:

$$IG = entropy_{prev} - \sum (entropy_{branch} * weight)$$

It is clear that larger *gains* are preferred.

Stopping Decision: Trees tend to overfit in the extreme case. In order to avoid this from happening, the dataset will be split into a training and validation set. A commonly used split is 70/30%. Where the validation set reaches its minimum, the procedure should be stopped.

> Decision trees can also be represented as sets of rules.

Decision trees model decision boundaries orthogonal to the axes.

Regression Trees

Decision trees can also be used to predict continuous targets. Impurity is measured by the Mean Squared Error. Low MSE means less impurity. Another way is to calculate the *F-statistic*. The assignment decision can be made by assigning the mean (or median) to each leaf node.

Neural Networks

A *Multi Layer Perceptron* is a neural network with one input, one hidden and one output layer. The hidden layer essentially works like a feature extractor by combining multiple inputs into features. The hidden layer uses a non-linear transformation function, the output layer uses a linear one.

Activation functions:

- Logistic
- Hyperbolic
- ReLU
- Linear

Typically fixed for each layer. For classification it is common practice to adopt logistic transformation in the output layer, since the outputs can in that case be interpreted as probabilities. Neural networks with one hidden layer are universal approximators, capable of approximating any function to any degree of accuracy.

Hyperparameters:

For NN's, the optimization is a lot more complex and the weights need to estimated using an iterative algorithm such as *Gradient Descent* or *Levenberg-Marquardt*. The cost function can have local minima, so iterative random weight initializations are necessary. Non-linear patterns require more hidden neurons.

Straightforward *iterative algorithm*:

1. Split the data into a training, validation and test set.
2. Vary the number of hidden neurons from 1 to 10 in steps of 1 or more.
3. Train a neural network
4. Measure its performance on the validation set
5. Choose the optimal number of neurons
6. Measure the performance on the independent test set

Hyperparameter tuning is *embarrassingly* parallel and can therefore be optimized faster using distributed techniques such as grid search with Functions As A Service.

Neural Networks are so powerful that they can even model the noise in the training set, which should be avoided. To prevent overfitting, weight regularization closely related to Lasso regression implemented by adding a weight size term to the cost function of the neural network.

Ensemble Methods

The aim of ensemble methods is to combine multiple analytical models into one more powerful model. Multiple models can cover different parts of the input space and as such complement each other's deficiencies.

Bagging

Bagging (bootstrap aggregating) starts by taking **B** bootstraps from the underlying sample. The idea is then to build a classifier (f.e. a decision trees, NN's, linear regression) for every bootstrap.

- With classification tasks, major voting schemes are used
- With regression tasks, the average outcome is taken

If perturbing the dataset by means of bootstrapping can alter the model construct, bagging will improve overall accuracy. For models robust with respect to the underlying dataset, it will *not* give much added value. It generally reduces variance and can be used to avoid overfitting, therefore constructing a model that *generalizes* better.

Boosting

Boosting works by estimating multiple models using a weighted sample of the data. Although there is *no single algorithm*. Boosting techniques will often iteratively reweight the data according to the classification error whereby misclassified cases get more attention. The technique works with either directly weighted observations, or with a dataset sampling according to the weight distribution.

The final ensemble is then a weighted combination (*strong learner*) from all of the individual models (*weak learners*). Multiple loss functions may be used to calculate the error misclassification rate is the most popular.

Boosting algorithms can be based on convex or non-convex optimization algorithms. Convex algorithms, such as AdaBoostand LogitBoost, can be "defeated" by random noise such that they can't learn basic and learnable combinations of weak hypotheses. Multiple authors demonstrated that boosting algorithms based on non-convex optimization, such as BrownBoost, can learn from noisy datasets and can specifically learn the underlying classifier of the Long–Servedio dataset.

For binary classification, the general algorithm is as follows:

1. Form a large set of simple features
2. Initialize weights for training images
3. FORALL T rounds
    1. Normalize the weights
    2. FORALL available features from the set: train a classifier using a single feature and evaluate the training error
    3. Choose the classifier with the lowest error

    4. Update the weights of the training images: increase if classified wrongly by this classifier, decrease if correctly
4. Form the final strong classifier as the linear combination of the T classifiers (coefficient larger if training error is small)

For multi-class classification:

The main flow of the algorithm is similar to the binary case. What is different is that a measure of the joint training error shall be defined in advance. During each iteration the algorithm chooses a classifier of a single feature (features that can be shared by more categories shall be encouraged). This can be done via converting multi-class classification into a binary one (a set of categories versus the rest), by introducing a penalty error from the categories that do not have the feature of the classifier.

A key advantage of Boosting is its simplicity. *A potential drawback* is that there may be risk of overfitting.

Random Forests

Random forests can be used with both classification and regression trees. Key in this approach is the dissimilarity amongst the base classifiers, which is obtained by adopting a bootstrapping procedure and the randomness of the splitting decisions. Diversity in base classifiers create an ensemble superior in performance.

1. Given a dataset with n observations and k inputs
2. m = constant
3. FOR t = 1 .. T
    1. Take a Boostrap sample of n observations
    2. Build a decision tree whereby *for each node of the three randomly choose m variables* on which to base the splitting decision
    3. Split on the best of this subset
    4. Full grow the tree *without pruning*

Evaluating Predictive Models

Evaluating Ensembles

Random Forests rank amongst the best performing models across a wide variety of prediction tasks and are perfectly capable of dealing with datasets that only have a few observations, but lots of variables. They are, however, black box models. One way to shed some light on the internal workings is by calculating variable importance.

Permutation (variable) importance:

1. Permute the values under consideration on the validation or test set
2. For each tree

$$VI(x_j) = \frac{1}{ntree} \sum_t error_t(D) - error_t(D'_J)$$

In a regression setting the error can be MSE whereas in classification it can be misclassification rate.

3. Order all variables according to their VI

Splitting up the dataset

When evaluating predictive models two key decisions need to be made:

1. The split up of the dataset
2. The performance metric

In the case of Neural Networks, the validation set is a separate sample since it is actively being used during the model development. A typical split up in this case is 40/30/30%.

In the case of small datasets, cross validation is often used, whereby the dataset is split up into K folds. A model is then trained on K - 1 training folds and tested on the remaining (validation) fold. In the extreme case, cross validation becomes *leave-one-out cross-validation*. One could let all models collaborate in an ensemble in order to acquire one *strong learner*.

For small datasets, one may adopt bootstrapping procedures. In bootstrapping, one takes samples with replacement from a dataset D.

We can approximate the performance as follows:

$$ErrorEstimate = 0.368 * Error(Training) + 0.632 * Error(Test)$$

Performance Measures for Classification Models

For continual regression, the scores need to be turned into predicted classes by adopting a cutoff score.

	Actual Negative	**Actual Positive**
Predicted Negative	TN	FN
Predicted Positive	FP	TP

$$ClassificationAccuracy = \frac{TP + TN}{TP + TN + FP + FN}$$
$$ClassificationError = \frac{FP + FN}{TP + TN + FP + FN}$$
$$Sensitivity = Recall = HitRate = TP/(TP + FN)$$
$$Specificity = TN/(TN + FP)$$
$$Precision = TP/(TP + FP)$$
$$FMeasure = \frac{2 * Precision * Recall}{Precision + Recall}$$

Integrating the Receiver Operating Characteristics yields the AUC which represents the probability that a randomly chosen *positive case* gets a higher score than a randomly chosen *negative case*.

A lift curve represents the cumulative percentage of churners per decile, divided by the overall population percentage of churners. Often summarized by top decile lift.

Other: Cumulative Accuracy Profile, Lorenz, Power Curve, Accuracy Ratio (= AUC*2 - 1)

Performance Measures for Regression Models

- Pearson correlation
- (R Squared) Coefficient of determination
- MAD
- MSE

Descriptive Analytics

> In descriptive analytics, often referred to as unsupervised learning, the aim is to describe patterns.

Association Rules

An association rule is an implication of the form X => Y. The rules measure correlational associations and should *not* be interpreted in a causal way!

Support and confidence are two key measures to quantify strength of an association rule. A frequent item set is an item set for which the support surpasses a certain preset threshold. The confidence measures the strength of an association and is defined as the conditional probability of the rule consequent, given the rule antecedent.

The lift, also referred to as the interestingness measure, takes into account the prior probability of the rule consequent. A lift value (smaller) larger than 1 indicates a negative (positive) dependence or substitution (complementary) effect.

Post processing rule mining by:

- Filtering out trivial rules
- Performing sensitivity analysis
- Use appropriate visualisation facilities
- Measuring economic impact

Whereas association rules are concerned with what items appear together at the same time, *sequence rules* are concerned about what items appear at different times.

Clustering

The aim of clustering is to split up a set of observations into clusters such that the homogeneity within each cluster is maximized and the heterogeneity between clusters is also maximized. Techniques are either hierarchical (agglomerative or divisive) or non-hierarchical (k-means). In order to decide on the merger or splitting, a distance measure is needed. This is often Euclidian or Manhattan. Various schemes are developed in order to define distance inbetween two clusters:

- Single Linkage

- Complete Linkage
- Average Linkage
- Centroid Method

Use a scree plot or dendogram to choose the optimal number of clusters.

Advantage: No prior number is specified

Disadvantage: Does not scale well

K-means clustering:

1. Select K observations as initial cluster centroids
2. Assign each observation to the cluster that has the closest centroid
3. When all observations are assigned, recalculate the position of the *k* centroids
4. Repeat until the centroid positions are stable

Self Organizing Maps:

Allows users to visualize and cluster high-dimensional data on a low-dimensional (rectangular/hexagonal grid) of neurons. A SOM is a feedforward neural network with two layers, whereby the output layer represents the grid. Each input is connected to all outputs. *Beware to standardize the data to zero mean and unit derivation first!* The neuron that is the most similar to a certain position is called the best matching unit.

A neighborhood function is derived.

Visualised using:

- A U (unified distance) matrix
- A component plane

Business Applications

Uplift Modeling

> Uplift modeling aims at estimating the net effect of a *treatment*. It allows users to optimize the selection of customers to include in marketing campaigns as well as further customization at the individual customer level of the campaign design. In essence, uplift modeling aims at shifting the focus towards a cluster of the customer base called the *persuadables*.

Response Modeling

Response modeling is used for setting up different types of marketing campaigns:

- Acquisition: Predicting prospects of interest
- Development: predicting interest prospects in additional goods/services
- Retention: predicting churn

Targeting a customer in a marketing campaign is referred to more generally as *treating* a customer. An action toward a customer is a *treatment*. We must know the *net effect* of a treatment on a customer rather than the *gross* effect to optimize the concrete actions that are undertaken. Uplift Modeling offers significant added value beyond the marketing context:

- Credit Risk Management
- Dynamic Pricing
- Biomedical clinical trial analysis
- Political Campaigning

Uplift modeling allows distilling the effects of these interactions from the data and accounting for the effects of interactions within the model. A key requirement is to distill the effects of behavior-interactions, is the *availability of the right data*. It is necessary to actively gather the required data by means of well-designed experiments or, alternatively, to passively gather the required data by tracking information on marketing campaigns.

Effects of a Treatment

We can distinguish 4 customer types:

	Y	**N**
N	Do Not Disturbs	Lost Causes
Y	Sure things	Persuadables

1. Sure Things: respond or purchase whether or not they are treated
2. Lost Causes: do not respond regardless of whether or not they are treated
3. Do Not Disturbs: adversely affected when treated, will respond otherwise
4. Persuadables: Respond solely when treated

Couple of notes:

- Sure things are costlier then Lost Causes (treatment cost + loss due lower pricing)
- A customer base may or may not consist of the above customer types

Data Pipeline

...