Predictive analytics implementation in the logistic industry

Yanka ALEKSANDROVA[1]

[1] University of Economics, Varna, Bulgaria
yalexandrova@ue-varna.bg

Abstract. Nowadays more than ever supply chain networks must meet increased demands. The digital transformation of the business is a necessity for the companies in the field of logistics in order to be able to increase their competitive advantages. During this transformation one of the most significant part plays the predictive analytics. Data driven decision making is crucial to supply chain activities. This however requires a more holistic view on implementation of predictive analytics in operational logistics processes. This paper aims to present possibilities for applying predictive methods in different operational processes in the context of a modern machine learning methodology framework and to demonstrate appropriate methods, techniques, algorithms, and software technologies. The scope of the research covers business processes in logistics organization. Several methodologies for design of predictive analytics frameworks have been evaluated and on this basis an adaptation of Microsoft Team Data Science Process is proposed. The methodology is demonstrated with an original practical implementation on a dataset provided by a logistics company. During each stage from the methodology suitable technologies, machine learning algorithms and evaluation measures have been applied. Conclusions are drawn regarding possibilities to implement the framework and to extract useful knowledge. Since the presented models are fitted to the used data set, the model explanation and interpretation is limited to the inherent data patterns and dependencies. Empirical results show that the best performing models are those trained with stacked ensembles and XGBoost algorithms. The model interpretation is implemented with SHAPley values and Partial Dependency Plots. The study is part of Project BG05M2OP001-1.002-0002-C02 "Digitalization of Economy in a Big Data Environment"

Key words: logistics, predictive analytics, machine learning, big data

1. Introduction

Companies in logistics industry are facing many challenges on their road to digitalization. They need not only to implement big data applications (Stoyanova, Vasilev, Cristescu, 2021) but apply appropriate methodological and technological tools for extracting useful knowledge from the rapidly growing data sources. Key technological trends that contribute and shape the logistics industry are also system integration, Internet of Things (IoT), cloud computing, blockchain, autonomous robots, virtual and augmented reality (Sulova, 2021).

The technological aspect of predictive analytics framework for the logistics industry has been a topic of several research publications (Mileva, Petrov, Yankov, Vasilev, Petrova, 2021). The purpose of the presented research, however, is to establish a methodical framework for the application of predictive analytics in logistics organizations. The framework covers all stages of application of prognostic analyses, starting from identifying business sense and selecting the data, choosing appropriate methods for predictive analysis, applying the extracted analytical knowledge to solve problems in different business processes implemented in logistics companies. The description of the framework has in addition to a theoretical-methodological basis and an applied aspect that examines and demonstrates appropriate information technologies.

2. Possibilities for implementing predictive analytics in logistics processes

To identify business analysis needs and the possibilities for the application of predictive analysis methods in the business area "Logistics", we explore the reference model SCOR (The Supply Chain Operations Reference). SCOR describes business activities related to all stages of meeting consumer demand and has a hierarchical structure. At the first level, the model is organized in 6 main management processes (APICS, 2020):

• **Plan –** this process describes the activities of developing a supply chain organization plan. This includes setting requirements, gathering information on available resources, balancing resources and requirements, and defining actions to provide the necessary resources in view of the defined requirements.

• **Source –** describes the activities related to orders, deliveries, receiving, transfer of materials, parts, products, and services. The scope of this process includes the preparation of delivery requests, supply planning, receiving and validation of supplies, accepting supply invoices, storage of goods.

• **Make –** the process describes the activities of conversion of materials into production. Activities include assembly, chemical treatment, maintenance, repair, recycling, production, renovation and other.

• **Deliver –** the process covers activities related to the creation, maintenance, and fulfillment of customer orders. Includes activities such as accepting customer orders, scheduling the orders execution, collecting, packaging, transporting, and issuing invoices.

• **Return –** the return process describes activities associated with the reverse flow of inventories. This process includes identification of stocks to be returned, determination of an appropriate return method, scheduling and dispatch.

• **Enable –** includes processes for maintaining and monitoring resources, relationships, assets, business rules, etc. to support the supply chain management. The scope of the process covers business rules management, performance management, data management, resource management, building and production capacity management, contract management, supply chain management, legal compliance management, risk management and supply chain supply. Graphically, the processes in the reference model are presented in fig. 1.

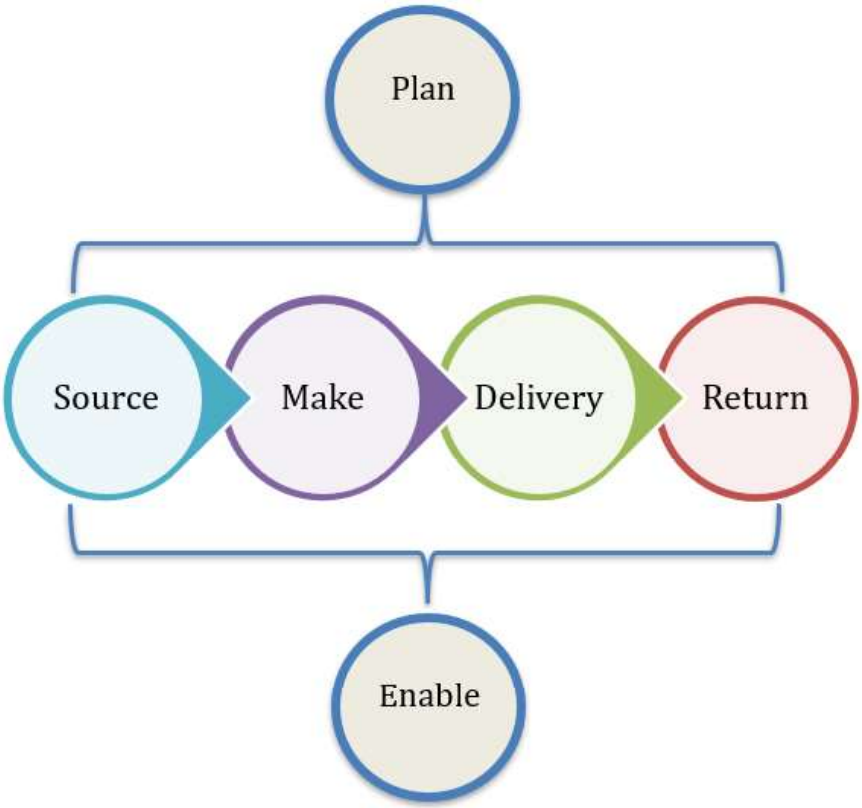

Figure 1. SCOR reference model
Source (APICS)

The structure of activities from “Plan” process makes it possible to identify several type activities. Type activities within level 3 processes (sPx.x) can be classified as follows:

- **Type 1**. Activities to identify, prioritize and aggregate resources.
- **Type 2**. Activities for identifying, prioritizing and aggregating requirements.
- **Type 3**. Balancing activities of requirements and resources.
- **Type 4**. Planning activities.

Predictive models can be successfully applied to type 1 and type 2 activities. Modelling should consider the level and scope of management decisions taken. Identification, prioritization and aggregation of resources and requirements can be implemented at the following levels:

- **Strategic.**
- **Tactical.**

- **Operational.**

The different levels of decision making determine the time horizon and the characteristics of the data used by prognostic models. In strategic-level decisions, data from longer periods of time are examined. Macroeconomic indicators, demographic trends, technological development forecasts, competitive intelligence, and others (Souza, 2014) are considered. The aim is to make long-term forecasts for the overall development of the supply chain. The prognostic models used at this level are mainly regression – auto-aggressive, multinominal, and logistic regression, as well as methods for predicting time series such as sliding average, regression sliding averages (ARMA) and auto-aggressive integrated moving averages (ARIMA), etc. (Liu et al., 2008), (Dhahri and Chabchoub, 2007; Patel et al., 2018).

When supporting tactical decision making, the study period is significantly shorter, usually within a year. Prognostic analyses are aimed at forecasting the products' demand and the necessary resources. Trend analysis methods are implemented as well as methods for detecting causal links between demand determining factors and target variables. In addition to traditional statistical methods, machine learning methods such as cluster analysis, market basket analysis (associative analysis), decision trees, neural networks, etc. are applied.

Operational solutions apply the same prognostic methods as those applied at tactical level, but with a more limited forecasting time horizon. Both prognostic methods and methods for real-time optimization of resources and requirements are used.

The "Source" process of the reference SCOR model describes planning and receiving supplies of inventories. The scope of this process includes the preparation of delivery orders, preparation of delivery schedules, receipt, verification of stocks on delivery, storage of stocks and processing of delivery factors. The Provisioning process consists of the following second-level processes:

- sS1 –Source Stocked Product
- sS2 –Source Make-to-Order Product
- sS3 –Source Engineer-to-Order Product

The subprocesses of sS1 and sS2 processes are similar, with only sS3 defined with two additional processes due to the specific characteristics of individual production. Third-level processes within the described three second-level processes can be presented in Fig. 2.

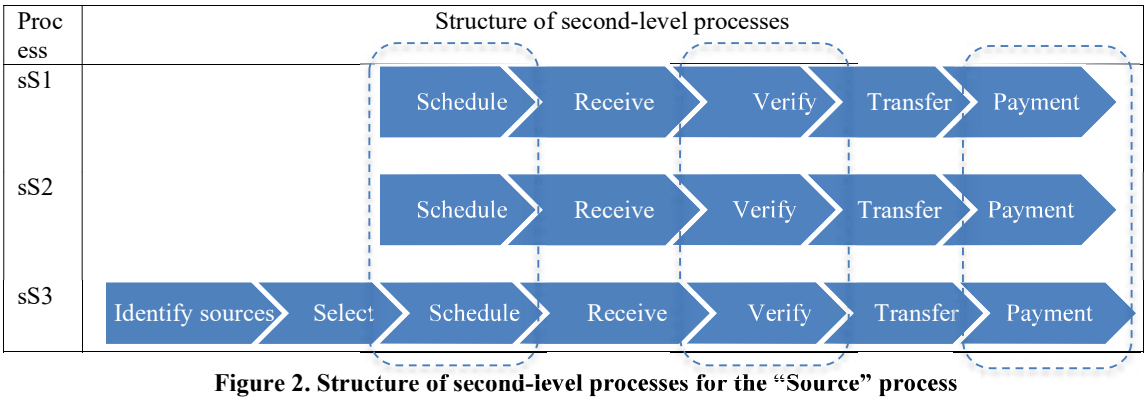

Figure 2. Structure of second-level processes for the “Source” process
Source: Own elaboration

The structure of the “Source” process reveals five type of third-level activities. They can be classified as follows:

- **Type 1.** Delivery planning processes.
- **Type 2.** Delivery processes.
- **Type 3.** Delivery inventory verification processes.
- **Type 4.** Processes for transferring products to storage warehouses.
- **Type 5.** Payment authorization processes.

With the greatest potential for applying prognostic models, in our view are type 1 ("Supply Planning") and type 4 ("Product Transfer"). In type 1 processes, appropriate prognostic models should be applied in the three second-level processes to forecast the necessary inventory levels and to draw up a supply schedule. However, this should also consider the specificities of the supply which are reflected in the selection of factor variables in prognostic modelling. When providing inventories in stock, the goal is to plan the optimal availability. The most important factor in building prognostic models here is the aggregated consumer demand in the relevant market. In this regard, prognostic models can focus on forecasting product demand and analyzing

the company's ability to influence and satisfy this demand. Significant factors to be considered in prognostic models in the process of securing stocks in stock also include macroeconomic indicators, seasonal fluctuations, mass production process parameters, production facilities, storage areas and equipment, etc. Since the process is aimed at maintaining optimal availability for products with wide use and mass production, regression time models can also be successfully applied.

When planning the schedule of inventory deliveries on customer orders, forecasting the optimal availability is more difficult to achieve, since the provision of inventories takes place after an accepted order from customers. In this regard, the main objective of prognostic models is to predict future orders, i.e., transactions within the framework of customer relationship management. This type of supply usually applies to products where the maintenance of stocks involves high costs, those not intended for mass use or, in the case of limited opportunities to predict the demand for the product type concerned.

The “Make” process of the reference SCOR model describes the activities related to the conversion of materials and the creation of finished products or services. This process consists of the following second-level processes:

- sM1 –Make-to-Stock.
- sM2 –Make-to-Order.
- sM3 –Engineer-to-Order.

The subprocesses of sM1, sM2 and sM3 processes are similar, with only sM3 adding an additional process in accordance with the characteristics of individual production. The third-level processes within the three second-level processes can be presented in fig. 3.

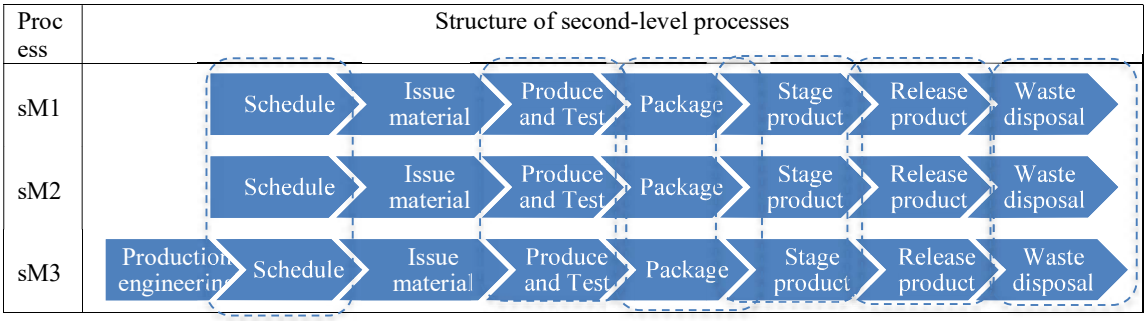

Figure 3. Structure of second-level process for the „Make“ process
Source: Own elaboration

Within the second-level processes, 6 type processes can be identified within the sM2 “Make” process. They can be systematized as follows:

- **Type 1.** Production schedule processes.
- **Type 2.** Processes for incorporating materials and resources into production.
- **Type 3.** Processes for creating new products and services.
- **Type 4.** Processes for packing the finished products.
- **Type 5.** Processes for revenue and storage of production.
- **Type 6.** Processes for picking finished products to ship from the warehouse.
- **Type 7.** Disposal processes, landfilling of waste.

Methods of prognostic analysis have their place in type 1 processes. When preparing a production schedule, the necessary quantities of finished products are predicted according to the results of the type 1 ("Planning") processes of the "Source" process. At the same time, in addition to prognostic methods, optimization methods should be applied to take account of the capacity of production capacity, seasonal fluctuations in load, accounting for planned and exceptional interruptions of the production process, forecasting and balancing all resources necessary for the implementation of the stages of the technological process.

In type 3 processes, prognostic methods may be aimed at predicting the quality of finished products, considering various factors influencing this. Appropriate data for the application of prognostic methods for predicting quality and preventing undesirable deviations from the production process schedule should be collected and analyzed in real time. For this purpose, intensive use of Internet of Things (IoT) technologies that allow recording, transmission, storage, and analysis of data related to the progress of the production stages is appropriate. Such data, except for the purposes of operational dimensions and production monitoring, may be used to apply prognostic methods for predicting key indicators of the performance of activities within the production cycle.

Prognostic methods can be also applied to type 6 processes. Within the subprocesses of this type according to the specifics of production – mass, on demand and individually – prognostic methods should be applied to forecast and plan material outflows. In mass production, product pick-up is linked to planned sales, which requires the use of sales forecasting results carried out in the planning process of the reference SCOR model to apply prognostic methods.

The proposed systematic approach for implementing predictive analytics in different types of processes within the SCOR reference model reveals certain opportunities for integrating extracted knowledge into the operational processes. The results from the applied approach could be used as a guidance to choose the appropriate predictive methods to support decision making and to facilitate the business process execution. However, in order to implement predictive analytics a more holistic view should be used which covers all the stages and steps in building and implementing an analytical application. This could be achieved by using a machine learning methodology framework covering all the activities from data acquisition to model acceptance.

3. Data acquisition

When applying machine learning methods for predictive analysis, specialists accept the following two assumptions:

1.There is a dependency between the result to be predicted and the factor variables that can be modelled and used for prediction.

2.The data collected are in the required quantity, structure, diversity, and quality so that this dependency can be revealed and modelled.

It is possible to establish during the application of prognostic analyses that either assumption is wrongly formulated, which in turn will lead to weak prognostic models. To avoid this, it is necessary to take a particularly careful approach to selecting appropriate data to detect alleged dependencies and interpret the business meaning of the data collected with a view to solving the prognostic tasks.

When selecting data, account should be taken of all possible internal and external sources. For their systematization, we propose to proceed from the logistics value chain offered by M.Porter (Porter, 1997). According to the value chain, two groups of data-generating activities can be identified, which can be used as sources for the purposes of the estimated dimension in the logistics subject area:

- **Main activities**
  - Inbound logistics – includes activities for supply of inventories, stock management, transport, production scheduling, quality control. Such types of activities are supported and automated by supply chain management systems and e-logistics platforms (Vasilev, 2017).
  - Operations – production, packaging, production control, quality management, maintenance.
  - Outbound logistics – output management, order management, shipping, customer delivery, invoicing.
  - Sales and marketing – customer management, sales management, promotions, sales analysis, marketing campaigns, research, and analysis.
  - Service – warranty and post-warranty service, maintenance, training, etc.
- **Additional activities**
  - Administration – accounting, finance management.
  - Human resources management – staff management, recruitment, training, planning of staff needs, health and safety environment, etc.
  - Technical and product development – product design, design of business and technological processes, product engineering, development.
  - Procurement – supplier management, financing, contract management and subcontractors, specifications, etc.

Experts should also consider external data sources such as data on competing companies, macroeconomic indicators for the specific region, market, sector, etc.

Data from internal and external sources are selected to solve the specific prognostic tasks and depend on the specifics of the subject area, the methods of analysis selected, the accessibility of the sources and the available IT infrastructure in the organization.

4. Data preparation and cleansing

The preliminary preparation of the data is a stage of utmost importance while applying the overall methodology for the development of a big data analysis application in logistics. The purpose of the data

preparation is to ensure the necessary quality and structure of the data set to be applied to the different methods of analysis.

Within the preliminary data preparation, the following more important stages can be defined:

1. Data cleansing and transformation
2. Feature engineering

Data quality management should be seen as an important strategy, inseparable from the overall process of implementing prognostic analyses. Data quality assurance can be implemented mainly in three environments:

- In the original data source system.
- During retrieving and loading the data when appropriate scripts are applied for processing the data at the time of their extraction from the various sources and loading them in the environment of the intelligent and analytical application.
- After loading the data in the data preparation area for analysis. In this case, the data shall be extracted and loaded without prior processing for quality assurance, and then the exploration and measurement of deviations from the set quality standards is carried out and techniques are taken to tackle the deficiencies in the data sets.

Companies may apply procedures to ensure the quality of the data in one, two or all the specified environments, depending on the specifics of the data sources used. In terms of efficiency and reusability with the greatest potential to ensure the necessary data quality is the implementation of the in the original data source system. This will achieve consistency and a single interpretation of data in both the source system and the analytical application. This is unfortunately not always possible, especially when data comes from external sources that lack the capacity to integrate data quality enhancing techniques into the system generating that data itself.

In any event, the causes of problems and poor data quality should be analyzed. These problems may be caused by (Loshin, 2013):

- different number of records that are retrieved and loaded in the analytical system. When working with large data, a huge number of records are worked on, and it is possible that during their processing some of the records may not be able to load in the preparation area.
- different level of detail (granulation) of data in different sources. A typical example of this is attributes such as name and address. In some sources, the name may be stored in three separate fields (name, surname, and last name), and in others, only in one field (Full Name).
- invalid values, especially in the case of codes from reference tables and nomenclatures.
- errors related to data transcription, such as spelling errors, use of abbreviations, replacement of Latin with Cyrillic, use of unrecognizable symbols, etc.

At the data cleansing and transformation stage, several techniques are applied, the most common being related to:

- Data conversion to the correct format.
- Identification and treatment of missing values. The high percentage of missing values is a serious obstacle to the application of prognostic methods, which must be overcome before the data can be analyzed. When examining for missing values, the cause of this problem should be identified first. Depending on the reasons, appropriate techniques for their elimination are then selected. The missing values may be missing completely at random (MCAR), missing at random (MAR), or missing not at random (MNAR) (Buuren, 2018). To solve problems with missing values, different methods can be applied, such as:
  - remove variables with a large percentage of missing values.
  - replace with medium or median (for numerical variables).
  - predicting missing values with methods such as linear regression, k-th Nearest Neighbor, random forest, etc.
- Outliers’ treatment. Outliers are such observations that differ significantly from others, which raises doubts about an unknown mechanism for the formation of these values (Hawkins, 1980). Most often, when determining outliers, the Tukey rule applies (Tukey 1977), according to which outliers are such values that are located at a distance from the median greater than 1,5 * Interquartile range (IQR). The presence of outliers in the data set may significantly affect the results of statistical methods and those of machine learning.

For the outliers’ treatment, the following approaches (Elshahawy, 2019), (Prabhakaran, 2017), (Sharma, 2018) are most often applied, such as:

- Elimination of observations with extreme values from the dataset. It should be considered that this loses information from other variables that do not have extreme values.
- Replacement of outliers with the mean or median. This is a rather simplified method of dealing with outliers and could lead to a deviation in models because the values of the other variables in the set are not considered. It should be applied with particular care in the presence of a significant number of outliers.
- Replacement of extreme values with variable values of 5th or 95th percentile values ("flooring" or "capping").
- Replacement of outliers with NA values (missing values) and predicting them with an appropriate method such as linear regression, kth Nearest Neighbor – kNN, etc.
- Grouping values into categories - variable discretization.

5. Modeling and implementation

When applying predictive methods, an appropriate methodology should be chosen. In this regard, we look at the most common methodologies applicable to predictive analysis. We believe that the application of a predictive analysis should be seen as a project for discovering and extracting knowledge and therefore in the survey overview are methodologies supporting the implementation of data mining processes.

In theory and practice, several methodologies are known for the realization of projects related to the discovery and extraction of knowledge. The most common are CRISP-DM (Cross-Industrial Standard Process for Data Mining), KDD (Knowledge Discovery in Database), SEMMA (Sample Evaluate Modify Model Assess) and Microsoft TDSP Lifecycle.

CRISP-DM

The CRISP-DM methodology was established in 1996-1997 by leading representatives of corporations with expertise in data mining – SPSS, Teradata, Daimler AG, NCR Corporation and OHRA. Crisp-DM's task is to standardize the process of knowledge discovery, structuring it in stages, defining the steps at each stage and defining the connections between the different stages.

CRISP-DM is described as a hierarchical process model containing a set of tasks on four levels of abstraction: stage, common task, specialized task, and process instance (Chapman, et al., 2000). The life-cycle stages of knowledge learning include business understanding, data understanding, data preparation, modeling, evaluation, deployment.

KDD

Knowledge Discovery in Database (KDD) as an iterative and interactive process that gained popularity in 1996 (Fayyad, Piatetsky-Shapiro, & Smith, 1996) and was based on earlier team research (Piatetsky-Shapiro & Frawley, 1991). KDD includes the following steps (Fayyad, Piatetsky-Shapiro, & Smith, 1996):

1. Study of the subject area and relevant previous knowledge, as well as defining the purpose of the process of knowledge extraction from the point of view of the business organization.
2. Selection of a data set suitable for achieving the objective.
3. Data cleansing and conversion.
4. Selection of important factors and eliminate unnecessary variables by dimensionality reduction.
5. Selection of appropriate group of data mining methods in accordance with the set objective.
6. Selection of appropriate algorithms to extract knowledge from the group of methods defined in the previous step and choosing their parameters.
7. Extracting knowledge from the cleansed and converted data set.
8. Interpretation of the results obtained from the previous step.
9. Using discovered knowledge.

The steps are organized in separate stages and implemented iteratively, and in the course of carrying out the entire knowledge-retrieval process it is possible to return to previous stages, similar to the interaction between the CRISP-DM stages.

SEMMA

The name SEMMA is an abbreviation of Sample, Explore, Modify, Model and Assess (SAS, 2020). SEMMA was established by the SAS Institute, supports all stages of the knowledge mining process, and is integrated as a logical organization of the functionality of SAS Enterprise Miner – one of the leading software platforms for knowledge mining. SEMMA organizes operators in SAS Enterprise Miner (so-called nodes peaks) in the mentioned five groups (tabs – tabs) corresponding to stages in the knowledge process.

Team Data Science Process

Team Data Science Process (TDSP) was created by Microsoft as a flexible, iterative methodology for building intelligent applications based on machine learning and artificial intelligence (Microsoft, 2020). TDSP is aimed at improving teamwork, sharing good practices, and using technology solutions from Microsoft or other leading IT companies. TDSP is organized as a life cycle and includes the following five iteratively performed stages (see fig. 4):

1. Defining the problem (Business understanding). The target variables and relevant metrics to measure the success of the project shall be determined. The relevant data sources shall be identified.
2. Data acquisition and understanding. Purified, quality datasets are created containing the necessary information on the target variables and related factors.
3. Modeling. The modeling phase is to determine the dependent and independent variables in the data set and to build an appropriate machine learning model.
4. Deploy. At this stage, the selected model is deployed into a production environment so that it can be presented for use by business users.
5. Customer Acceptance. On this stage the implemented model is validated by the customers, assessing the fulfillment of the goals set and meeting the user requirements.

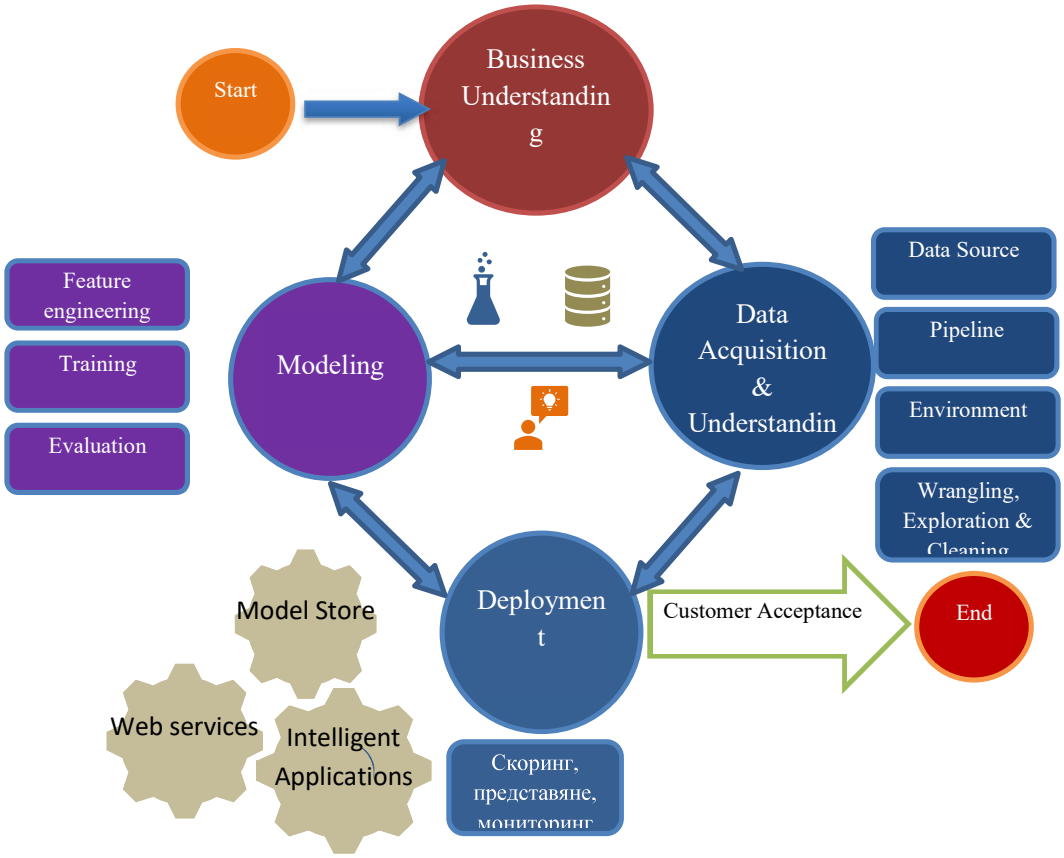

Figure 4. The Data Science Project
Source: Microsoft

For each of the stages, Microsoft offers full support for the activities performed through various technological solutions, documentation, guides, good practices, etc. Although TDSM has been demonstrated with Microsoft technologies and platforms such as Azure Machine Learning, Azure SQL Server, SQL Server, etc., the company also offers ready-made free solutions for two of the most used machine learning programming languages– R and Python.

Comparison of the methodologies presented

The study of the structure and content of the stages of the methodologies examined shows quite similarities and overlaps in some stages. The stages that occur in all four methodologies under consideration are:

understanding and researching data, transformation, modeling, and evaluation of models. All methodologies suggest iterative implementation of the stages with the possibility of returning to a previous stage to reflect changes.

CRISP-DM, KDD, and SEMMA do not include a data collection stage. They assume that the data is already available and, after defining the business problem, proceed to their retrieval. TDSP looks at data collection as part of a second stage, Data Acquisition and Understanding. Different ways of collecting data in the context of the development of large data and ubiquitous connectivity have been considered. TDSP looks at technologies to build applications for automatic data collection and entry in different environments such as Azure Blob Storage, SQL Azure Databases, SQL Server on Virtual Azure Machines, HDInsight (Hadoop), Synapse Analytics, and Azure Machine Learning. Data mining is seen as an integral part of the overall knowledge retrieval and application project.

Finally, a serious advantage of TDSP over CRISP-DM, KDD and SEMMA is the inclusion of the Customer Acceptance stage. As stated, at this stage the implemented model is presented to end users for validation. The level of implementation of the objectives set shall be assessed before moving to regular operation.

Since its establishment in 1996, CRISP-DM has rapidly gained popularity and over the years has been considered the most used methodology (Piatetsky, 2014) and de-facto standard for the knowledge extraction process (Mariscal , Marban, & Fernandez, 2010), (Putler & Krider, 2012), (Wiemer, Drowatzky, & Steffen, 2019). However, CRISP-DM 2.0 was never created, and the official website http://crisp-dm.org/ long ago seized to function. Currently, CRISP-DM is supported by IBM as part of the SPSS–Modeler analytical and statistical software, but without updating. There are also such problems with KDD, which has not been developed since its creation and has not been adapted to new fast-growing information technologies such as big data, the Internet of Things, real-time analytics, etc. SEMMA has a more limited scope and is strongly tied to the SAS Enterprise Miner analytical platform.

The comparative analysis of CRISP-DM, KDD, SEMMA and TDSP provides grounds to conclude that TDSP is the most complex methodology. It covers the whole knowledge process and can be applied to the construction of data mining and machine learning applications. TDSP is constantly evolving and adapting to dynamic changes in the field of big data, artificial intelligence, machine learning, the Internet of Things, etc. Due to these advantages, we believe that TDSP is best suited for application when applying predictive analysis methods in the field of logistics.

6. Model evaluation

The assessment of trained machine learning models is an important milestone in the overall methodology for the construction and application of analytical models. The specific methods, indicators and evaluation scales depend on the type of tasks solved. It is recommended to use multiple indicators and evaluation techniques when assessing the performance of models to gain a fuller picture of the prognostic capabilities of models and their presentation on new data unknown to them.

In supervised training, the model-building experts must strike a balance between adapting to the learning set and generalization. Adapting to the learning set is the goal of learning, with the algorithm aiming to identify as much as possible all dependencies and patterns in the set. The uncontrolled learning however may lead to over-fitting of the model, reducing its ability to generalize. To achieve balance and avoid overfitting, different techniques are applied for the regularization while training, which depend on the selected machine learning algorithm.

Evaluation metrics are presented as a number from 0 to 1 or a percentage of 0% to 100%. They should be considered together as they reflect differently the presentation of the model. The sensitivity shows how good the model is for detecting positive cases, and specificity reflects the ability to identify negative cases. Comparing models based on only one metric is incorrect, especially when the data set is strongly unbalanced. Accuracy indicator can only be considered when the distribution of the two classes in the data set is approximately the same (50:50). In other cases, when it comes to unbalanced datasets, it is more correct to report Balanced Accuracy, which is calculated as an arithmetic mean of specificity and specificity.

ROC (Receiver Operating Characteristics) curve is also used for graphical representation of the model's prognostic strength compared to a random classification model (Ghavami, 2019), (Thanaki, 2018), (Kuhn M. , 2013, p. 265). Since the ROC curve does not depend on the threshold chosen, it can be used as a reliable measure of model performance and when comparing two or more models. The metric, which is calculated based on the built ROC curve, is an area under the curve (AUC). The ideal model has an area under the curve = 1. This is the model that generates only correct predictions. The random model, which resembles guessing predictions, has an

AUC = 0.5. Its curve coincides with the diagonal of the graph. A classification model is the better as its area under the curve is closer to 1.

7. Model explanation and interpretation

The application of the extracted knowledge should be preceded by an analysis of the performance of the trained models to assess their prognostic strength. Evaluation metrics depend on the machine learning methods selected. At the same time, however, it is also essential to apply appropriate means, technologies and models for interpreting and explaining models.

The ability of a model to be explained and understandable to consumers is associated with the concept of explainability and interpretability. The two concepts have a large degree of overlap and are often used as replaceable (Gall, 2018), (Molnar, 2020), (Miller, 2019). The interpretation of the model determines how much a person can understand the reasons for a prediction (Miller, 2019). The ability of a model to be interpreted is associated with the extent to which people can systematically predict and consider the outcome of its application (Been, Khanna, & Koyejo, 2016). The higher the degree of interpretation of a model, the easier it is to understand the reasons and mechanisms for making a decision or generating a particular prognosis. The explainability of machine learning models, on the other hand, determines the possibility of presenting the results of the models in an intelligible form to consumers (Gall, 2018), (Molnar, 2020).

The ability to interpret the model increases confidence in its application (Ribeiro, Singh, & Guestrin, 2016), (Lipton, 2017), (Was, Khanna, & Koyejo, 2016), (Babel, Buehler, Pivonka, Richardson, & Waldron, 2019), If consumers are not convinced of the legality of the model, they will not use it. Confidence in the model can be expressed in confidence in the correctness of the predictions generated for a particular case and confidence in the performance of the model.

Some of the machine learning models, such as logistical regression and single classifiers such as classification and regression trees, have an inherent ability to be explained (Norton, Dowd, & Maciejewski, 2019). Decision trees are another example of a model with built-in interpretation. From the graphical representation of the tree, rules based on each variable's contribution to the result can be easily generated and explained. However, the problem with this type of model is that they often show poor performance on the test set. There is a counter-proportional relationship between the built-in interpretation and prognostic capability of the models, i.e., the easier it is to explain the model with its inherent characteristics, the greater the deviations of the generated predictions (Bussmann et al., 2020).

Interpretation and explanation of black box models, lacking built-in possibilities for explanation, is realized by creating an agnostic model to explain the generated predictions from the model. The models are agnostic as they do not depend on the algorithm used to train the model to be interpreted. This allows to explain each model, regardless of its degree of complexity and the algorithm used.

The interpretation and explanation of the model can be considered in two directions – globally and locally. The first explains the overall mechanism of operation of the model, brings out global dependencies and templates by aggregating all observations. The local level of explanation makes it possible to interpret and justify individual forecasts by showing the influence of factor variables on the generation of the specific prediction for the observed case.

Different methods are used to explain the global model, but the most applied are variable importance and partial dependence plots. In cases where the model does not have a built-in interpretation, it is necessary to build a new model that can explain the mechanism of operation and the logic of generating individual predictions. This new surrogate model, in turn, must be sufficiently simple and easily interpretable. To solve such problems, different methods are applied, and one of the most efficient and currently used is SHAP (Shapley Additive exPlanations) proposed by (Lundberg et al., 2019).

SHAP values are based on Lloyd Shapley's work (Shapley, 1953) in the field of game theory and explain the prognosis through the marginal contribution of each factor variable. The values of independent variables are seen as coalition participants, and Shapley values show what their contribution is to the "prize" – the specific prediction that is generated by the model.

Based on SHAPley, the Lundberg and Lee values (Lundberg et al., 2019) look at each explanation of the predictions generated by a model as a model and refer to it as an "explanation model." They define a class of additive factor contributing models that include various methods such as: linear additive model, LIME (Local Interpretable Model-agnostic Explanations) (Ribeiro, Singh, & Guestrin, 2016), DeepLift (Deep Learning Important FeaTures) (Shrikumar, Greenside and Kundaje, 2017), SHAPley values, etc.

8. Predictive analytics framework implementation

The possibilities for applying prognostic analyses will be demonstrated with an experiment in the environment of the h2o platform. H2O is a scalable distributed, fast, primarily memory-based open-source platform for machine learning and predictive analytics. It enables the construction of machine learning models on big data and provides easy implementation of models in a working environment. The platform is built and provided by the H2O.ai company, whose corporate mission is democratization of artificial intelligence. In the latest research for 2020 by the consulting company Gartner, H2O.ai is listed as a visionary in the field of data science and machine learning platforms (Krensky, etc., 2020) and cloud services for the development of artificial intelligence (Baker, Elliot, Sicular, Mullen, & Brethenoux, 2020).

The basic code of H2O is written in Java. The platform uses distributed key/value repository to access and retrieve data, models, objects, etc. between all nodes and machines. The algorithms are implemented on a distributed Map/Reduce platform and use the Java Fork/Join multi-reduction platform. The data is read in parallel, distributed between clusters and stored in ram in column compressed format.

Through REST application interface is provided access by external programs or JSON scripts through http protocol. A Web interface (Flow UI) is available for easy construction of machine learning models, as well as an interface for access to the platform from the R (H2O-R) (LeDell, et al., 2020) and Python (H2O-Python) environment.

The h2o.automl() function can be used to automate the machine learning process, including automatic learning and optimization within user-defined time limits. In addition, ensembles models are also built – Stacked Ensemble of the best algorithms and Stacked Ensemble from the best algorithms of a family. AutoML also displays a leaderboard table that compares the performance of the best models against the training and validation data set.

The algorithms that are applied to train models using h2o.automl() function are:

- Distributed Random Forest (DRF) – includes Random Forest and Extremely Randomized Trees (XRT).
- Generalized Linear Model (GLM).
- XGBoost Gradient Boosting Machine (XGB).
- Gradient Boosting Machine (GBM).
- Deep Learning (DL).
- StackedEnsemble.

Stacked Ensemble is a heterogeneous ensemble algorithm that finds the optimal combination of a set of prognostic algorithms using a process called stacking (H2O.ai, 2020). These ensemble models support regression, binary and multinominal classification. The groundbreaking scientific study demonstrating theoretically the effectiveness of combining models and stacking them into an ensemble model was published in 2007 (van der Laan, Polley, & Hubbard, 2007), and further developed in 2010 (Polley & van der Laan, 2010). These two publications use the term "Super Learner" to mean heterogeneous ensemble models with the arrangement of models based on different algorithms and the use of cross-validation to build the combining algorithm, the so-called "super learner".

To train the models in h2o we use the sample data set "DataCo Supply Chain Dataset" (Constante, F. & Silva, F. (2019)) The data are provided by the company DataCo – a leader in the field of data management and information management solutions. The set contains 180519 observations representing orders from customers in different regions of the world. The data structure allows different machine learning algorithms to be applied. In this case, we will use binominal classification algorithms to predict the risk of delay in order delivery.

To solve the task, the relevant factor and result variables are selected. A description of the variables in the selected subset is given in Table 1. The target variable Late_delivery_risk has two possible values - 1 (there is a risk) and 0 (no risk).

With the help of function h2o.automl() 20 basic classification and 2 heterogeneous ensemble models are trained. The maximum training time is set to 2 hours, with the logloss metric controlling its end. The assessment of the performance of models on validating multitudes is presented in Table 2. During the training, 2 heterogeneous ensemble models, 7 models with GBM algorithm, 7 XGBoost, 3 Deep Learning and one from DRF, XRT and GLM are built.

Table 1

Models’ performance (sorted by AUC)

model_id	auc	logloss	aucpr	mean_per_class_error
StackedEnsemble_AllModels_AutoML	0.8978	0.4018	0.9216	0.2008
StackedEnsemble_BestOfFamily_AutoML	0.8944	0.4071	0.9190	0.2060
GBM_4_AutoML	0.8937	0.4041	0.9188	0.2020
GBM_5_AutoML	0.8936	0.4052	0.9187	0.2012
GBM_grid__1_AutoML_model_2	0.8830	0.4238	0.9107	0.2174
GBM_3_AutoML	0.8802	0.4261	0.9093	0.2212
DRF_1_AutoML	0.8738	0.4487	0.9039	0.2292
GBM_2_AutoML	0.8660	0.4470	0.8992	0.2409
GBM_1_AutoML	0.8456	0.4750	0.8849	0.2659
XRT_1_AutoML	0.7820	0.5442	0.8400	0.3629
GBM_grid__1_AutoML_model_1	0.7787	0.5376	0.8406	0.3751
XGBoost_grid__1_AutoML_model_3	0.7584	0.5497	0.8264	0.4171
GLM_1_AutoML	0.7583	0.5601	0.8231	0.4075
XGBoost_1_AutoML	0.7543	0.5532	0.8233	0.4114
XGBoost_grid__1_AutoML_model_4	0.7534	0.5581	0.8218	0.4000
XGBoost_grid__1_AutoML_model_1	0.7525	0.5540	0.8220	0.4251
XGBoost_3_AutoML	0.7503	0.5537	0.8210	0.4225
DeepLearning_1_AutoML	0.7503	0.5557	0.8236	0.4479
XGBoost_2_AutoML	0.7501	0.5606	0.8196	0.4366
XGBoost_grid__1_AutoML_model_2	0.7473	0.5560	0.8188	0.4198
DeepLearning_grid__1_AutoML_model_1	0.7455	0.5719	0.8191	0.4432
DeepLearning_grid__2_AutoML_model_1	0.7324	0.5671	0.8108	0.4780

Source: Own calculations

As evident from table 2 the highest AUC values are for the two heterogeneous ensemble classification models – StackedEnsemble_AllModels and StackedEnsemble_BestOfFamily. After them are arranged gradient boosting models (GBM). The significance of the predictions generated in the best model (StackedEnsemble_AllModels) is shown by graphical representation of metalearner in Fig. 5.

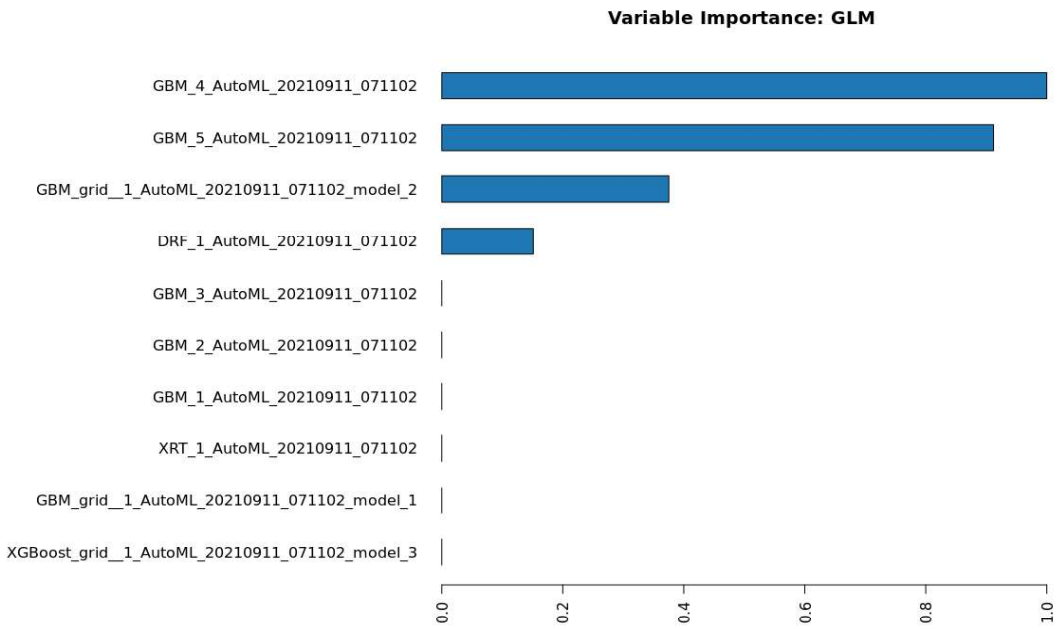

Figure 5. Model importance in the metalearner
Source: Plot generated in h2o

The best model is applied to the test set. The model has an accuracy of 0.8962 and a balanced accuracy of 0.8940. The distribution of the target class is approximately equal, with 45.2 % of the cases being of Class 0 (accepted as positive) and 54.8 % being of Class 1 (negative). When dividing the data set into a learning and test set, this ratio is maintained.

Model interpretation at the global level includes exploring the influence each variable has on the target. It is well illustrated with the variable importance chart (see Fig. 6). According to the chart the most influential independent variable is “Shipping Mode”, followed by “Customer City”, “Order State” and “Order Country”. The relative importance however doesn’t show the direction of the influence and the influence of each categorical value from mentioned variables.

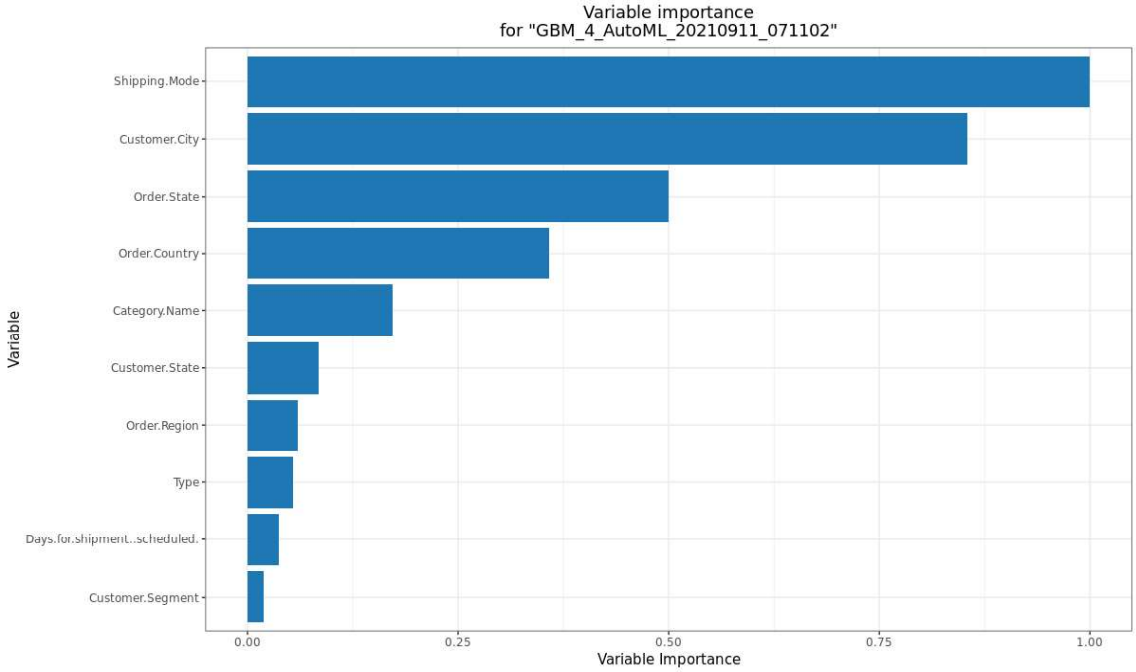

Figure 6. Variable importance plot
Source: Plot generated in h2o

A summary of the contribution of each variable for each instance from the test dataset is shown on fig. 7. The SHAP summary plot is generated using the best base model – GBM_4. This chart can be used to interpret the influence of different features on the target variable. As shown on the diagram the most influential variable is the shipping mode, followed by customer city, order type, order country and scheduled days for shipment. Results from the SHAP summary plot could be used by experts to identify the most important factors and to explore their impact on the target variable. SHAP contribution could be also used as a mean for determing possibilities to improve business processes by mitigating the negative impact of the explored variables.

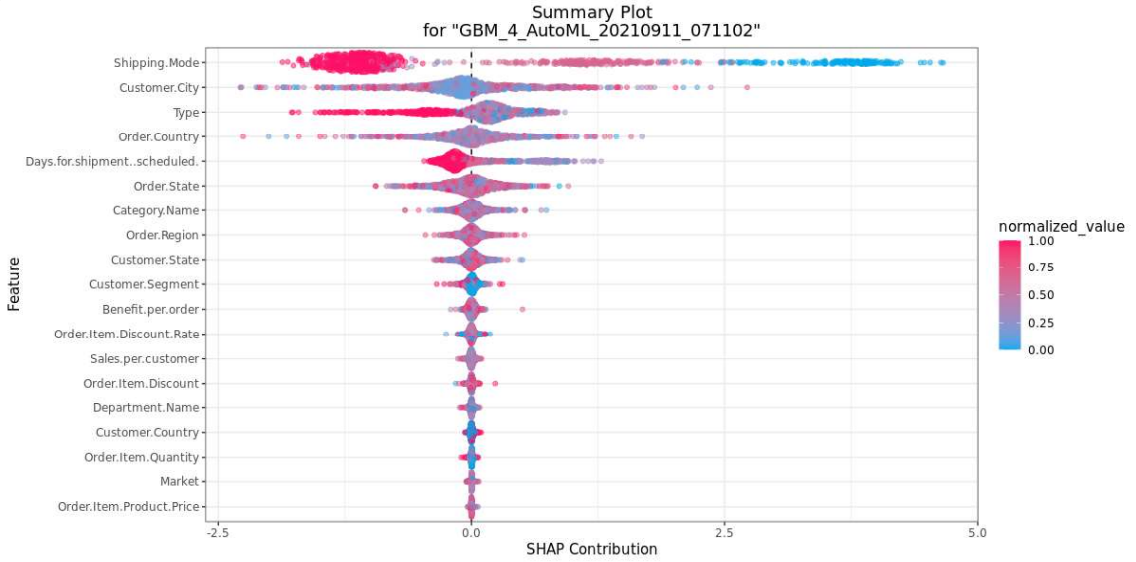

Figure 7. SHAP summary plot
Source: Plot generated in h2o

Another useful tool for interpreting heterogeneous ensemble models is the partial dependence plot (PDP). Such plot is shown on figure X for the most important variable – the categorical variable shipping mode.

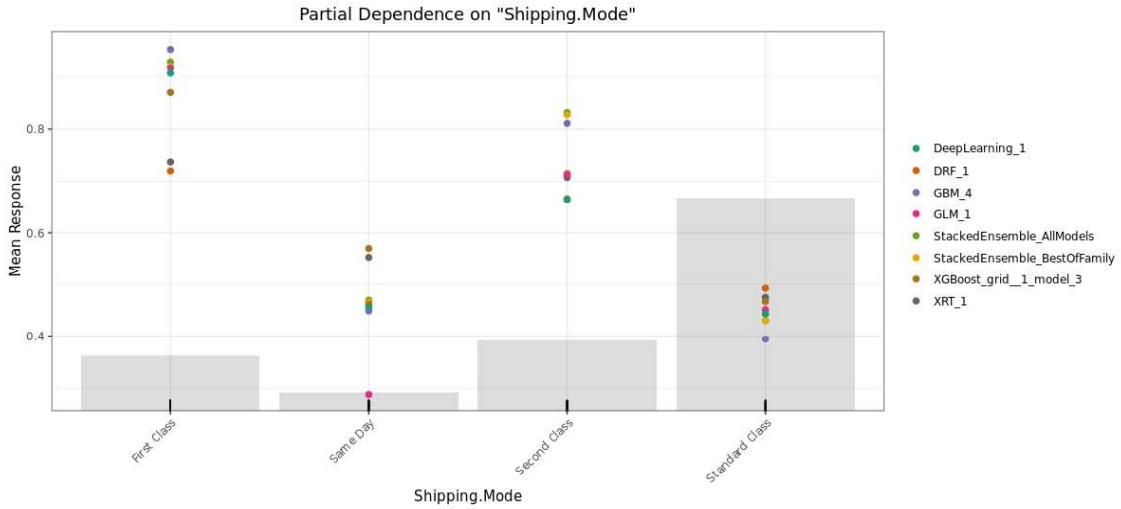

Figure 8. Partial Dependency Plot for Shipping.Mode variable
Source: Plot generated in h2o

As evident from Fig. 8, most base models show consent on the direction and degree of influence of different values of shipping mode. For example, the greatest mean response, e.g., greater risk of delay, is associated with first- and second-class shipping mode. A far lower delay risk is associated with standard and same day shipping mode. The partial dependence plot of shipping mode variable could point out areas for additional exploration to identify possibilities for reducing the delay risk and thus improving the overall customer order processing.

9. Conclusion

The demonstrated implementation of predictive analytics for the logistic industry outlines the stages, methods and technologies that could be used to integrate the extracted knowledge into operational logistic processes as defined in the SCOR model. This paper reveals the importance of model evaluation, comparison, and explanation with appropriate measures, models, and methods. Results from the presented study can be used as a guidance for applying a predictive framework in organizations from the logistic industry.

Literature

APICS. (2020). SCOR framework. Available at: http://www.apics.org/apics-for-business/frameworks/scor

Apley, D., & Zhu, J. (2016). Visualizing the Effects of Predictor Variables in Black Box Su-pervised Learning Models. arXiv:1612.08468 [stat.ME]. Available at: http://arxiv.org/abs/1612.08468

Babel, B., Buehler, K., Pivonka, A., Richardson, B., & Waldron, D. (2019). Derisking machine learning and artificial intelligence. Available at: https://www.mckinsey.com/business-functions/risk/our-insights/derisking-machine-learning-and-artificial-intelligence [Accessed 17.06.2020]

Baker, V., Elliot, B., Sicular, S., Mullen, A., & Brethenoux, E. (2020). Magic Quadrant for Cloud AI Developer Services. Available at: https://www.gartner.com/doc/reprints?id=1-1YGJKJ5P&ct=200224&st=sb

Been, K., Khanna, R., & Koyejo, O. (2016). Examples are not enough, learn to criticize! Criticism for interpretability. *29th Conference on Neural Information Processing Systems (NIPS 2016)*, 29, pp 2280-2288. Barcelona, Spain.

Burk, S., & Miner, G. (2020). It's All Analytics! The Foundations of AI, Big Data, and Data Science Landscape for Professionals in Healthcare, *Business, and Government*. CRC Press.

Buuren, S. (2018). Flexible imputation of missing data (2 ed). Chapman & Hall/CRC.

Buuren, S., & Groothuis-Oudshoorn, K. (2011). mice: Multivariate Imputation by Chained Equations in R. *Journal of Statistical Software*, 45(3). doi:10.18637/jss.v045.i03

Chapman, P., Clinton, J., Kerber, R., Khabaza, T., Reinartz, T., Shearer, C., & Wirth, R. (2000). CRISP - DM 1.0. Available at: https://www.the-modeling-agency.com/crisp-dm.pdf

Elshahawy, S. (2019). The Ultimate R-Guide to process missing or outliers in dataset. Available at: https://medium.com/: https://blog.usejournal.com/the-ultimate-r-guide-to-process-missing-or-outliers-in-dataset-65e2e59625c1

Gall, R. (10 2018 r.). Machine Learning Explainability vs Interpretability: Two concepts that could help restore trust in AI. Available at: https://www.kdnuggets.com/2018/12/machine-learning-explainability-interpretability-ai.html

Ghavami, P. (2019). Big Data Analytics Methods: Analytics Techniques in Data Mining, Deep Learning and Natural Language Processing. Walter de Gruyter GmbH & Co KG.

H2O.ai. (2020). Stacked Ensembles. Available at: https://docs.h2o.ai/h2o/latest-stable/h2o-docs/data-science/stacked-ensembles.html

Hawkins, D. (1980). Identification of Outliers. Springer.

Krensky, P., den Hamer, P., Brethenoux, E., Hare, J., Idoine, C., Linden, A., Choudhary, F. (11 2 2020 r.). Magic Quadrant for Data Science and Machine Learning Platforms. Available at: https://www.gartner.com/doc/reprints?id=1-1YGJKJ5L&ct=200224&st=sb&mkt_tok=eyJpIjoiWkRKaFpqZzFOV0kzWTJSbSIsInQiOiJiVjdtVW52dVB5aEJCNEJ0UnVZWUpDVFg0Zk9yY1BHdm5ubkM1N1B1bkVaeGFPTlwvd1FiczFNWkRSOTVhVmhaSWtHRHV4OEphT1owcXRJeGJqM2F4RnFpeFBNeXVYV0xVbE5vVHlza1wv

Landis, R., & Koch, G. (1977). The Measurement of Observer Agreement for Categorical Data. *Biometrics*, 33(1), pp 159-174. doi:10.2307/2529310

Larose, D., & Chantal, L. (2015). Data Mining and Predictive Analytics (2 ed). Wiley.

LeDell, E., Gill, N., Aiello, S., Fu, A., Carno, A., Click, C., Malohlava, M. (2020). h2o: R Interface for the 'H2O' Scalable Machine Learning Platform. R package version 3.32.0.1. Available at: https://github.com/h2oai/h2o-3

Lipton, Z. (2017). The Mythos of Model Interpretability. arXiv e-prints. Available at: arXiv:1606.03490 [cs.LG]

Loshin, D. (2013). Business Intelligence: The Savvy Managers Guide (2 ed). Morgan Kaufmann. doi:https://doi.org/10.1016/C2010-0-67240-3

Mariscal , G., Marban, O., & Fernandez, C. (2010). A survey of data mining and knowledge discovery process models and methodologies. *The Knowledge Engineering Review*, 137–166. doi:10.1017/S0269888910000032

Mayor, E. (2015). Learning Predictive Analytics with R. PACKT Publishing.

McHugh, M. (2012). Interrater reliability: the kappa statistic. *Biochemia Medica*, 22(3), pp 276-282.

Microsoft. (2020). Team Data Science Process Documentation. Available at: https://docs.microsoft.com/en-us/azure/machine-learning/team-data-science-process/

Mileva, L., Petrov, P., Yankov, P., Vasilev, J., Petrova, S. (2021). Prototype model for big data predictive analysis in logistics area with Apache Kudu. *Electronic journal* “*Economics and computer science”*, Issue 1, Varna

Miller, T. (2019). Explanation in Artificial Intelligence: Insights from the Social Sciences. *Artificial Intelligence*, 267, pp 1-38. doi:10.1016/j.artint.2018.07.007

Mohri, M., Rostamizadeh, A., & Talwalkar, A. (2018). Foundations of Machine Learning. The MIT Press.

Norton, E., Dowd, B., & Maciejewski, M. (2019). Marginal Effects—Quantifying the Effect of Changes in Risk Factors in Logistic Regression Models. *JAMA*, 321(13), 1304–1305. doi:10.1001/jama.2019.1954

Peter, B., & Andrew, B. (2017). Practical Statistics for Data Scientist. O'Reilly Media.

Piatetsky, G. (2014). CRISP-DM, still the top methodology for analytics, data mining, or data science projects. Available at: KDnuggets: https://www.kdnuggets.com/2014/10/crisp-dm-top-methodology-analytics-data-mining-data-science-projects.html

Polley, E., & van der Laan, M. (2010). Super Learner in Prediction. *U.C. Berkeley Division of Biostatistics Working Paper Series*. Available at: http://biostats.bepress.com/ucbbiostat/paper266

Porter, M. (1997). Perspectives on Strategy. Springer. doi:https://doi.org/10.1007/978-1-4615-6179-8

Putler, D., & Krider, R. (2012). Customer and Business Analytics: Applied Data Mining for Business Decision Making Using R. Routledge.

Ribeiro, M., Singh, S., & Guestrin, C. (2016). “Why Should I Trust You?” Explaining the Predictions of Any Classifier. *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Demonstrations* (pp 97-101). San Diego, CA: Association for Computational Linguistics. doi:10.18653/v1/N16-3020

Samuel, A. (2000). Some studies in machine learning using the game of checkers. *IBM Journal of Research and Development*, 44(1.2), pp 206-226. doi:10.1147/rd.441.0206

SAS Institute. (2020). Predictive Analytics. Available at: https://www.sas.com/en_us/insights/analytics/predictive-analytics.html

Shah, A. (2016). Machine Learning vs Statistics. Available at: https://www.kdnuggets.com/2016/11/machine-learning-vs-statistics.html

Shalev-Shwartz, & Ben-David, S. (2014). Understanding Machine Learning. From Theory to Algorithms. Cambridge University Press.

Shapley, L. (1953). A value for N-person games. (H. W. Tucker) *Contributions to The Theory of Games*, 2(28), pp 307-317.

Sharma, N. (2018). Ways to Detect and Remove the Outliers. Available at: https://towardsdatascience.com/ways-to-detect-and-remove-the-outliers-404d16608dba

Shmueli, B. (18 12 2019 r.). Multi-Class Metrics Made Simple, Part III: the Kappa Score (aka Cohen’s Kappa Coefficient). Available at: https://towardsdatascience.com/multi-class-metrics-made-simple-the-kappa-score-aka-cohens-kappa-coefficient-bdea137af09c

Souza, G. (2014). Supply Chain Analytics. *Business Horizons*, 57, 595-605.

Stehman, S. (1997). Selecting and interpreting measures of thematic classification accuracy. *Remote Sensing of Environment*, 62(1), pp 77-89. doi:https://doi.org/10.1016/S0034-4257(97)00083-7

Stoyanova, M., Vasilev, J., Cristescu, M. 2021. Big Data in Property Management. Applications of Mathematics in Engineering and Economics: *Proceedings of the 46th Conference on Applications of Mathematics in Engineering and Economics (AMEE '20),* AIP Conference Proceedings 2333.

Sulova, S. (2021). Big data processing in the logistics industry, *Electronic journal* “*Economics and computer science”*, Issue 1, Varna

Thanaki, J. (2018). Machine Learning Solutions: Expert techniques to tackle complex machine learning problems using Python. Packt Publishing Ltd.

Tukey, J. (1977). Exploratory Data Analysis. Addison Wesley.

van der Laan , M., Polley, E., & Hubbard, A. (2007). Super Learner. *Statistical Applications in Genetics and Molecular Biology*, 6(1). doi:10.2202/1544-6115.1309

Wiemer, H., Drowatzky, L., & Steffen, I. (2019). Applications (DMME)—A Holistic Extension to the CRISP-DM Model. *Applied Science*, 9(12). doi: //doi.org/10.3390/app9122407

Zhang, A. (2017). Data Analytics. Practical Guide to Leveraging the Power of Algorithms, Data Science, Data Mining, Statistics, Big Data, and Predictive Analysis to Improve Business, Work, and Life. CreateSpace Independent Publishing Platform.

Vasilev, J., 2017. E-logistics in the context of globalization. (in Bulgarian) Varna: Publishing house "Science and Economy".