1 Good Practice Profile*

Title	Predicting the tariff code of a material master using artificial intelligence
Summary	Assigning tariff numbers to a product is a time-consuming manual procedure. For foreign trade, each company must classify its products as a precondition for export/import processes. Bosch has established a shared service of highly qualified experts to classify the Bosch products with the correct tariff numbers. To support this unit, Bosch applied supervised machine learning algorithms to train models to predict tariff numbers with a high accuracy. The enabler for this innovative solution is the combination of the high material master data quality and the availability of a standardized global tariff code classification process at Bosch. As a result of this innovative solution, quality, speed, and accuracy of the shared classification service can now be significantly improved.

Profile	
Contact	Hinselmann, Georg Flegl, Valk Pfreundner, Klaus
Contact Job Title	
Company name	Robert Bosch GmbH
Industry	Automotive
Company Overview	The Bosch Group is a leading global supplier of technology and services. It employs roughly 402,000 associates worldwide (as of December 31, 2017). The company generated sales of 78.1 billion euros in 2017. Its operations are divided into four business sectors: Mobility Solutions, Industrial Technology, Consumer Goods, and Energy and Building Technology. As a leading IoT company, Bosch offers innovative solutions for smart homes, smart cities, connected mobility, and connected manufacturing. It uses its expertise in sensor technology, software, and services, as well as its own IoT cloud, to offer its customers connected, cross-domain solutions from a single source. The Bosch Group's strategic objective is to deliver innovations for a connected life. Bosch improves quality of life worldwide with products and services that are innovative and spark enthusiasm. In short, Bosch creates technology that is "Invented for life." The Bosch Group comprises Robert Bosch GmbH and its roughly 440 subsidiary and regional companies in 60 countries. Including sales and service partners, Bosch's global manufacturing, engineering, and sales network covers nearly every country in the world. The basis for the company's future growth is its innovative strength. At 125 locations across the globe, Bosch employs some 64,500 associates in research and development.

2 Motivation and goals

- *Summarize the initial situation and the surrounding conditions before the start of the project*
- *Describe the drivers and goals for this project*
- *Describe the planned innovation of this project*

The Bosch portfolio contains millions of individual products. Of outmost importance for foreign trade is the assignment of a correct customs tariff number to the product, the so called commodity code. This code is a binding legal requirement and the precondition for customs clearance. Consequences of a wrong code can be significant, e.g. a delayed customs clearance or severe penalties.

Due to its criticality, the assignment of commodity codes is implemented at Bosch as global uniform process. The classification process is triggered as soon as a material with a defined maturity is created or extended to an ERP system. The change is captured on a central material master data system and forwarded to SAP Global Trade Services (GTS), where the classification request item is entered in a classification worklist. A central team of highly qualified experts (Center of Expertise; CoE) assigns the correct codes on GTS within hours.

After the labor-intensive manual classification process, the commodity codes are distributed to the local ERP systems.

As of mid-2018, we were facing the following situation.

1. There were about 200.000 classified material masters in the system, with about 200 different commodity codes for 26 countries and the European Union.
2. The global classification process has been live for two years. With an increased maturity level of the data, classification becomes a repetitive task for the experts. Many classification tasks are similar to previous requests. However, the work still has to be completed, requiring expert knowledge.
3. Currently about 70% of the Bosch business units are covered by this global classification process. A significant increase in number of classification requests is expected until end of 2018. However, the CoE service is limited in capacity.

Triggered by CDQ knowledge exchange we identified the above scenario as an ideal use case for machine learning to support the CoE.

Our approach learns the relationship between the material master data of the Bosch products and the assigned commodity code. The resulting model is used to predict the commodity code based on the master data. Our solution is based on supervised machine learning.

3 Approach

- *Describe the approach to design and implement this project*
- *Describe the change impact on organization and systems*
- *Describe the major challenges your company had to overcome*

In a supervised machine learning approach, a self-learning algorithm learns the relationship between a set of features and a label. The features in our case are represented by the fields of the material master, and the label is the national/international commodity code. The result of the training process is a model, a mathematical decision function.

Below is an example of a commodity code. The first six digits (harmonized tariff system number) represent the international classification, the remaining digits describe national requirements.

- 85: Portable electric lamps
  - 8512: Electrical lighting or signalling equipment
    - 851290: Parts
      - 8512900000: Parts of equipments of heading No.85.12; 品目8512所列装置的零件;指车辆等用照明,信号装置,风挡刮水器,除霜器等零

We began with a set of fields which are maintained at the time point when a classification process is initiated (Figure 1). The majority of features were categorical, followed by text features and numerical features. We also observed that for the several fields the data was noisy (i.e. dummy values, first item in the selection list) and incomplete.

Feature	Meaning
Material Number Range	The first four digits of the material number range is centrally governed at Bosch. It is related to a business unit and a group of products.
Material Type	Material type (e.g. “sample”, “semi-finished”, “saleable material”, etc.)
Industry Sector	Type of industry (e.g. pharmaceutical, engineering, etc.)
Material Group	Classification of material (e.g. Thermo technology, Camera systems)
Material Standard Description (English)	Describes the material in textual format. Concatenated text from various sources. Available in multiple languages.
Term Code	Standardized textual description of the material (68.000 unique term codes, with multiple translations)
Product Hierarchy	Product hierarchy (e.g. Starter Motors -> Starter Motors 1.1 kW.
Type Short Description	Textual description of the material, less details than Material Standard description.
Base Unit of Measure	Base unit of measure (e.g. “G”, “KG”)
Net Weight	Net weight of the material (e.g. “2.0”)

Figure 1: Features used for building a training set

We extracted two data sets to train the supervised learning algorithm: one for international codes, and a second with national codes for China.

A Random Forest Classifier was used to train the models with 11.000 and 50.000 unique instances, both having about 200 different labels. Other classifiers were evaluated as well, and had either a higher computation time (e.g. Support Vector Machines) or had a significantly worse predictive performance (e.g. Decision Trees). The Random Forest Classifier algorithm computes an ensemble of decision trees from a training set. It then predicts the label of an unknown sample as the consensus vote of the majority of the trees. We used the implementation provided by scikit learn. We divided the two data sets into two stratified random splits: 90% for training the model and a 10% test set to evaluate the accuracy. The training sets were used for parameter optimization and feature selection.

The Random Forest Classifier can also be used to calculate the feature importance (Figure 2). The feature importance is the average relative importance of the feature in the set of decision trees. The feature relevance is weighted by the node (weight is determined by the number of samples explained in the training set). This approach can be used to eliminate the features which are either irrelevant or redundant in our training data. Features with a low importance are either correlated with other features or they are simply useless for the training process (e.g. dummy values, constant values, empty values). We focus on the most importance features: Material Standard Description, Term Code, Product Class, and Material Number Range.

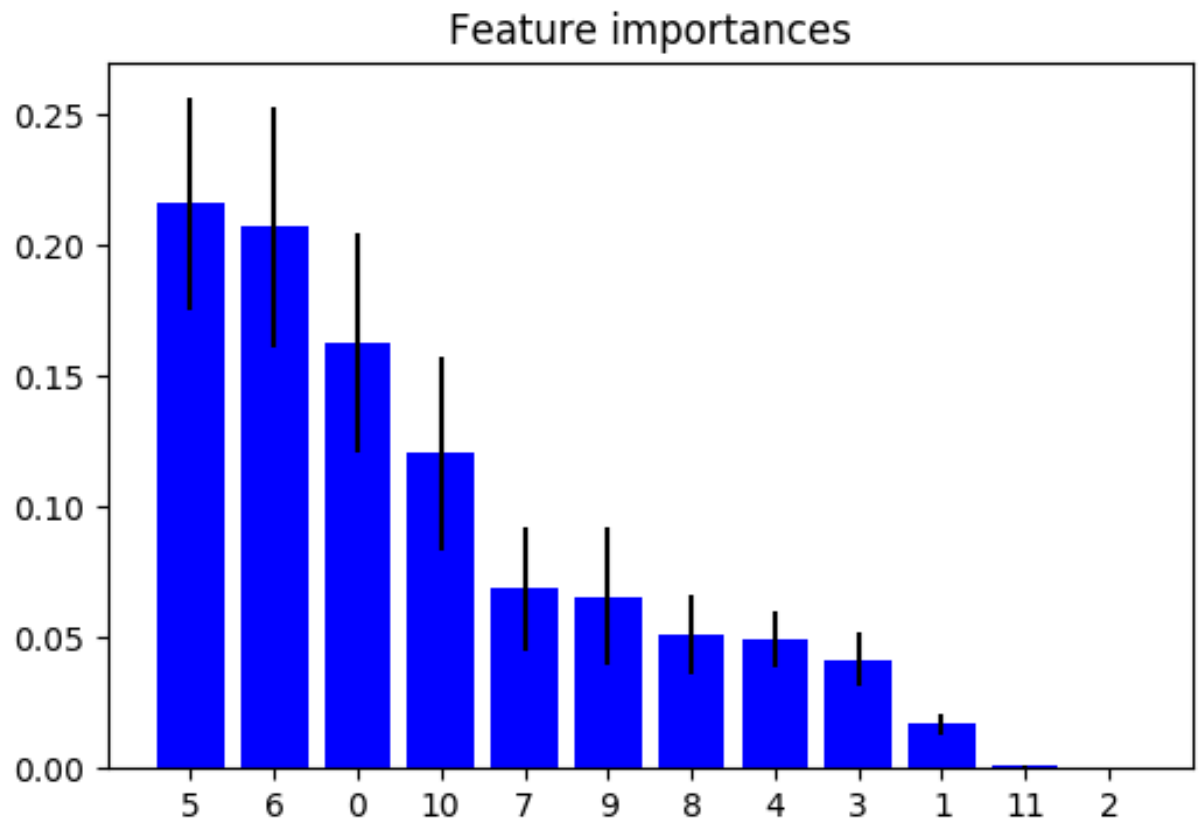


Figure 2: Feature importance

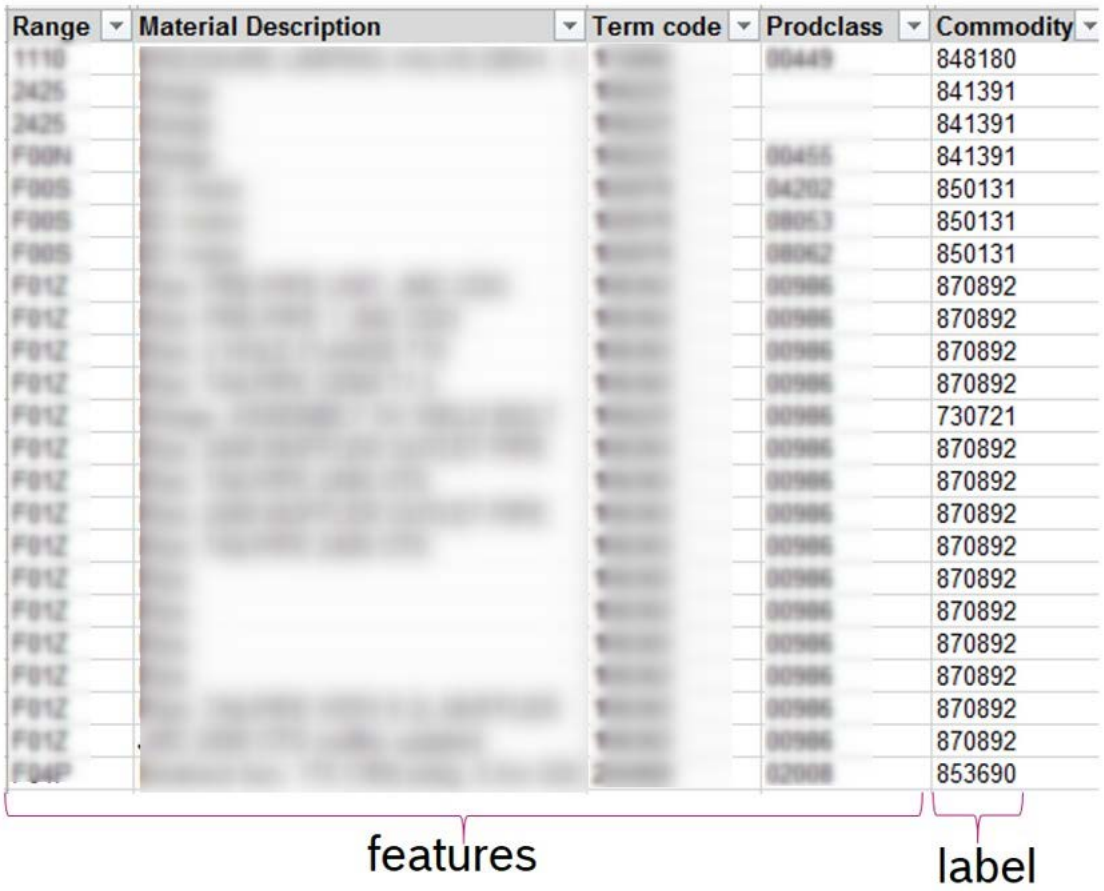


Figure 3: Final set of master data attributes (features) and – in this case – the international commodity code (label).

The Material Standard Description was mapped to a bit vector. All single words (1-gram) and combination of two words (2-grams) were extracted in the given sequence. The 1,2-grams were mapped to a 200 bit vector using their hash codes as seeds for a random number generator which produces a bit in [0, 199] for each code. This approach has two main advantages

1) Exact match are not required: "CONTROL UNIT 10V" and "CONTROL UNIT" share two of three 1-grams, but are not equal.
2) It reduces the runtime of the Random forest classifier where the number of dimensions in a multiplier in the runtime complexity. An advantage of this approach is that not all tokens have to be kept in memory.

Bringing all pieces together, we have an artificial intelligence (as depicted in Figure 4) that is able to solve our classification problem. The model has four input parameters: Material Standard Description, Term Code, Product Class, and Material Number Range.

The text is internally encoded as 200 bit vector. In total, the model is based on 200+3 features. The output is a commodity code plus a confidence score for the prediction.

To validate the accuracy of the prediction we measured the performance on an external data set. For 50 trees with a maximum allowed depth of 25, the model was calculated in less than a minute on both training sets. Prediction is done in a matter of milliseconds.

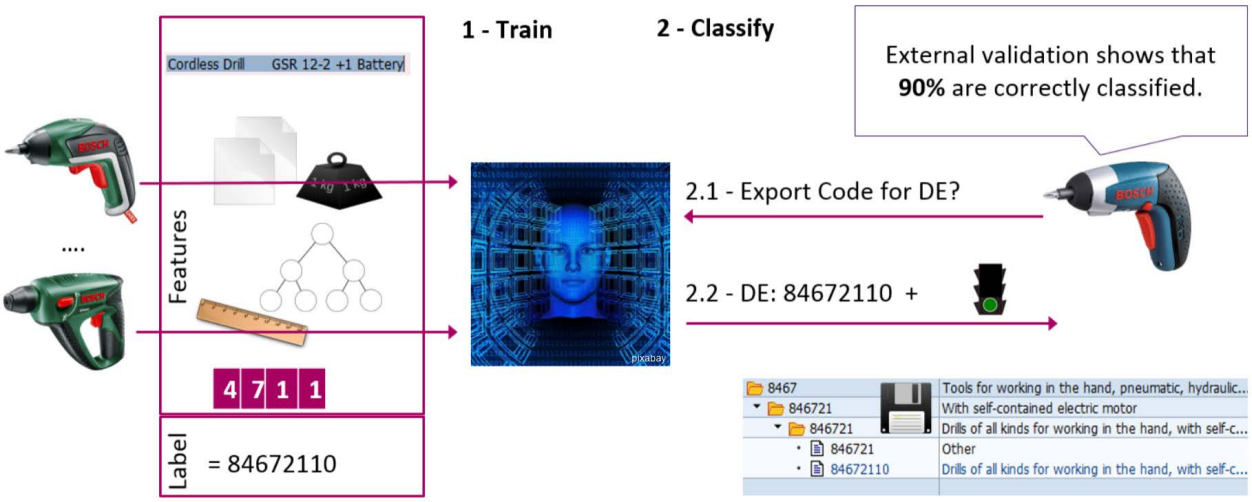


Figure 4: Model for the prediction of the commodity codes

The measured external accuracies were shown on Table 1: Results of external validation on external test set.

Table 1: Results of external validation on external test set

Benchmark	Accuracy
International Codes	90,2%
National Codes (China)	89,6%

In a second use case we corrected wrong international commodity codes in the system with the model trained on the full data set. The corresponding records were labeled with a dummy code. Examples of the results are shown in Table 2: Examples of corrected codes by the model.

Table 2: Examples of corrected codes by the model

Parameters	Prediction Result
• Material Text = "EL CONTROL UNIT AIRBAG 8" • Term Code = 123456 • Material Range = 0123 • Product Class = 00123	• Confidence = 77,7% • [85] ['Electrical machinery and equipment and parts thereof; sound recorders and reproducers, television image and sound recorders and reproducers, and parts and accessories of such articles'] o [8537] ['Boards, panels, consoles, desks, cabinets and other bases, equipped with two or more apparatus of heading 85.35 or 85.36, for electric control or the distribution of electricity, including those incorporating instruments or apparatus of Chapter 90, and numerical control apparatus, other than switching apparatus of heading 85.17.'] ▪ [853710] ['For a voltage not exceeding 1,000 V']
• Material Text = "FLAT MOTOR-AND-GEAR ASSY" • Term Code = 7890123 • Material Range = 4567 • Product Class = 045678	• Confidence = 100,0% • [85] ['Electrical machinery and equipment and parts thereof; sound recorders and reproducers, television image and sound recorders and reproducers, and parts and accessories of such articles'] o [8501] ['Electric motors and generators (excluding generating sets).'] ▪ [850110] ['Motors of an output not exceeding 37.5 W']
• Material Text = "Valve Set" • Term Code = 4567890 • Material Range = 8901 • Product Class = 09012	• Confidence = 93,4% • [84] ['Nuclear reactors, boilers, machinery and mechanical appliances; parts thereof'] o [8409] ['Parts suitable for use solely or principally with the engines of heading 84.07 or 84.08.'] ▪ [840999] ['Other: Other']

New classifications are added to the training data and the new domain knowledge is included in the model after the next training iteration.

The CoE is now supported by an artificial intelligence model to solve routine cases faster. Based on the confidence score, they can easily see if the classification by the model is pure chance or if this is within the applicability domain of the model. If the confidence score is low the CoE experts need to do the classification manually. The final decision about the code remains with the expert.

▶ AI Service Integration

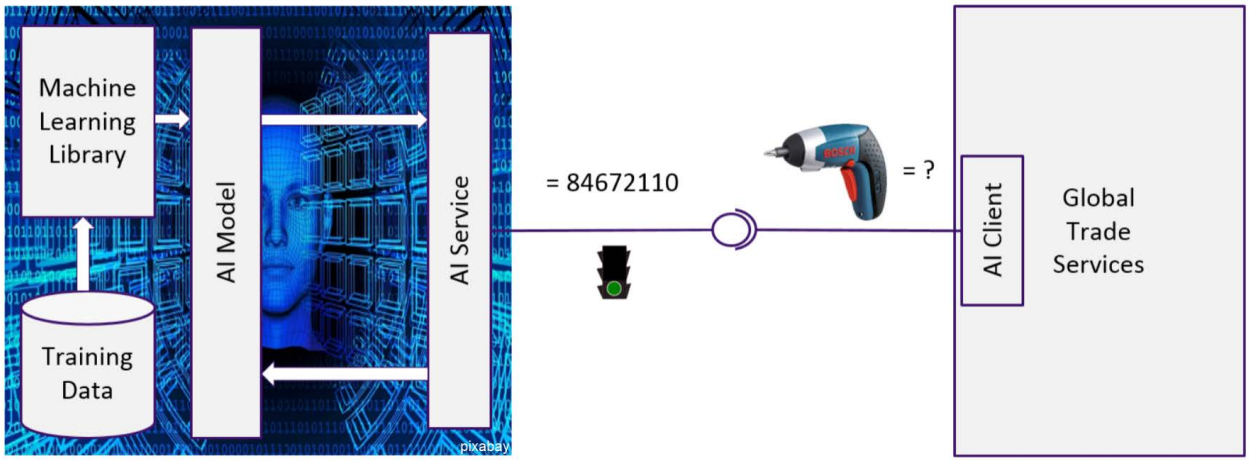
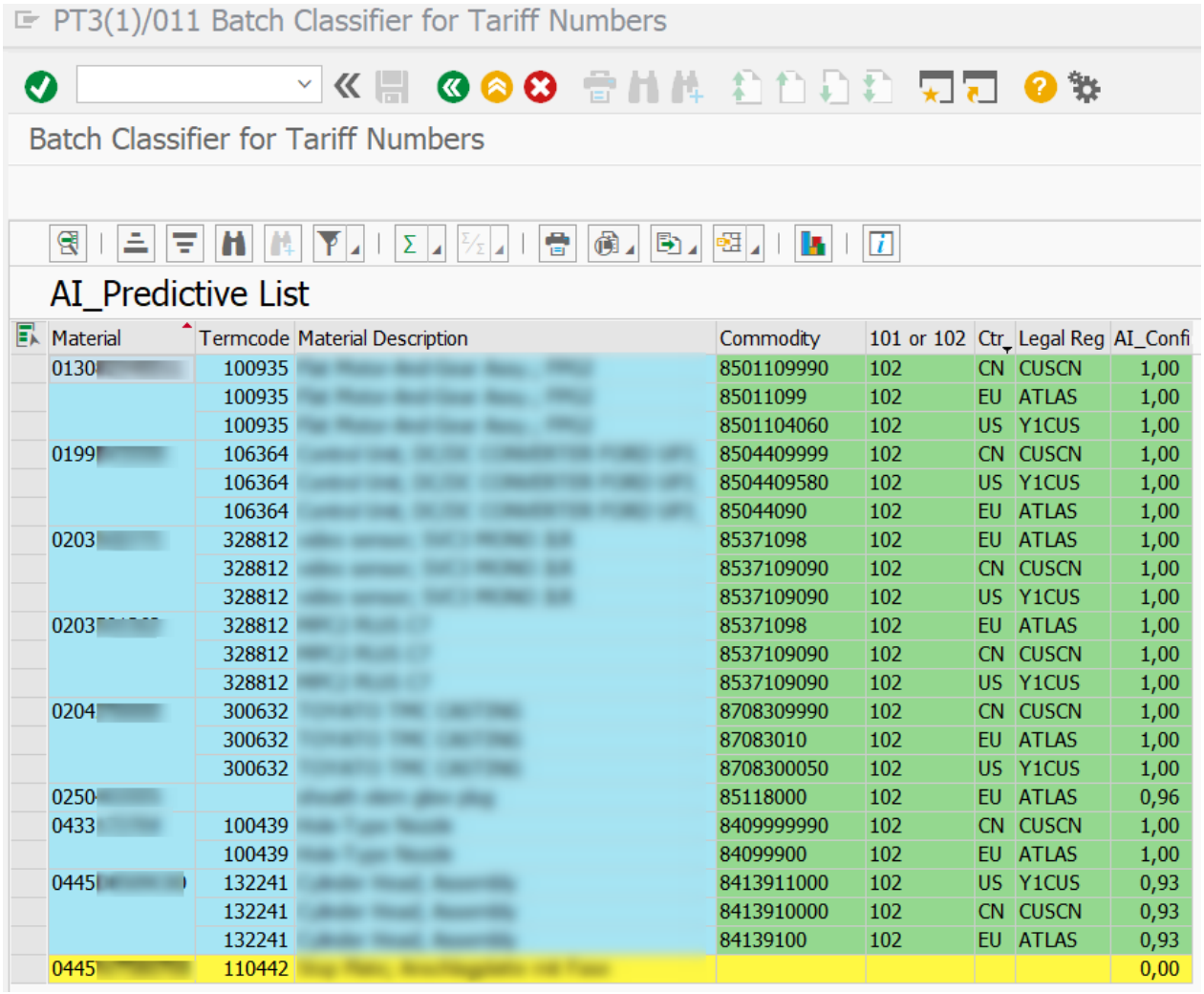


Figure 5: AI Service System integration



Material	Termcode	Material Description	Commodity	101 or 102	Ctr	Legal Reg	AI_Confi
0130	100935		8501109990	102	CN	CUSCN	1,00
	100935		85011099	102	EU	ATLAS	1,00
	100935		8501104060	102	US	Y1CUS	1,00
0199	106364		8504409999	102	CN	CUSCN	1,00
	106364		8504409580	102	US	Y1CUS	1,00
	106364		85044090	102	EU	ATLAS	1,00
0203	328812		85371098	102	EU	ATLAS	1,00
	328812		8537109090	102	CN	CUSCN	1,00
	328812		8537109090	102	US	Y1CUS	1,00
0203	328812		85371098	102	EU	ATLAS	1,00
	328812		8537109090	102	CN	CUSCN	1,00
	328812		8537109090	102	US	Y1CUS	1,00
0204	300632		8708309990	102	CN	CUSCN	1,00
	300632		87083010	102	EU	ATLAS	1,00
	300632		8708300050	102	US	Y1CUS	1,00
0250			85118000	102	EU	ATLAS	0,96
0433	100439		8409999990	102	CN	CUSCN	1,00
	100439		84099900	102	EU	ATLAS	1,00
0445	132241		8413911000	102	US	Y1CUS	0,93
	132241		8413910000	102	CN	CUSCN	0,93
	132241		84139100	102	EU	ATLAS	0,93
0445	110442						0,00

Figure 6: Result List Classifier: Original input, predicted tariff codes, and confidence score. Last row is below threshold. Filtered for US, CN, and EU country codes.

4 Self-Assessment

- *Describe the achieved results of this project considering*
  - *Data excellence*
  - *Business value*
  - *Innovation*
- *State the planned next steps (if applicable)*
- *Summarize the lessons learned*

Data excellence

- The solution enables an automated assignment of commodity codes with high accuracy (90%)
- The solution assists classification experts with suggestions of commodity code and reviews commodity

- The pillars of the solution are the established master data processes, the central assignment of the commodity codes, and a modern machine learning approach.
- The solution is not limited to a specific domain knowledge, the model works for other areas (e.g. power drills and control units) if covered by training data.

Business value

- Our approach accelerates the classification process by providing decision support with a confidence score for the prediction.
- Solution provides an intelligent duplicate check.
- The model improves the quality of the expert decision by reviewing the classifications of the experts.
- The approach is highly scalable.

Innovation

- Artificial intelligence analyzes the master data and reveals the most relevant fields. This enables the master data organization to focus on the data quality of these specific attributes.
- The model can be considered as kind of swarm intelligence combining and harmonizing the knowledge of many experts in one robust tool.
- The first use case in the Bosch ERP landscape where AI supports the end users.

Planned next steps

The model was successfully implemented in a Proof of Concept and successfully tested by the business experts.

Next steps:

- Further evaluation of architecture and suitable platforms (e.g. R Enterprise Server and HANA Predictive Analytics)
- Further integration of the solution into standard business processes.

Summarize the lessons learned

- First lessons learned so far:
    - The master data area provides many use cases for machine learning approaches with real business benefits
    - Today's hardware and machine learning technology are able to handle the Bosch business cases
    - A good and innovative approach with fantastic results
- Further experiences of the productive daily usage can be shared in one of our CC-CDQ workshops