HECTO
HP Engineering Classifier Tool

Project Overview

Sasquatch Engineering

Agenda

01 **Introduction.**

02 **Project Overview**

03 **Data Overview**

04 **Modeling**

05 **Project Results**

06 **Conclusion**

Problem/Opportunity statement

The manual classification of hundreds of new components into UNSPSC (United Nations Standard Products and Services Code) categories each month presents HP with a significant business challenge.

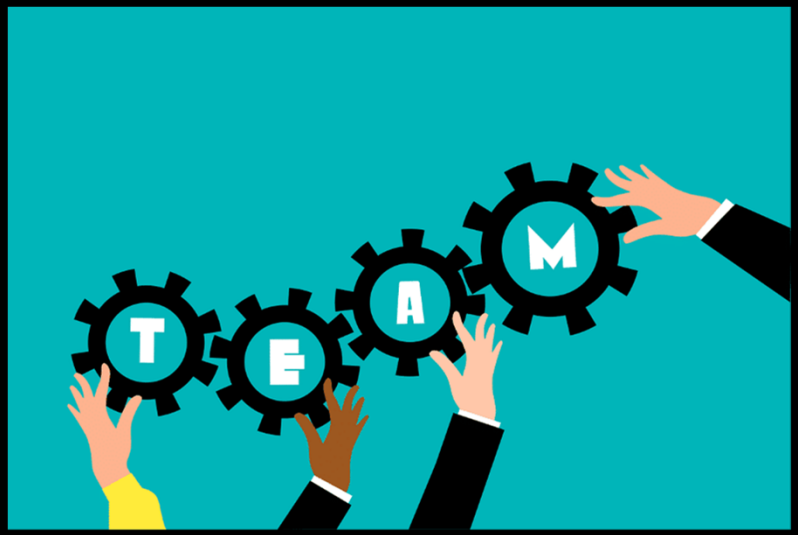
Current State :

The current process, reliant on engineers, is time-consuming, prone to human error, lacks scalability, and struggles with the growing volume of products. This inefficiency hampers productivity and leaves challenging parts, requiring detailed specifications, to be handled manually.

Opportunity :

The project aims to create a more efficient and automated solution to streamline UNSPSC assignments,reduce costs and improve efficiency in the process.

About Us



Sasquatch Engineering is a dedicated team of data scientists, machine learning engineers, and industry experts with a deep understanding of your industry's unique challenges and opportunities. We recognize the enormous potential of machine learning to transform your business operations and deliver tangible results for our clients. The Sasquatch engineering consulting team’s members are:

- Reed Ballesteros
- Manojkumar Damodaran Pillai
- Shawn Tay
- Nan Li

Project Objectives

Automate Classification Process:
- *Objective*: Develop a machine learning model to automate the classification of products into their respective UNSPSC categories using product information.
- *Rationale*: Streamlining the classification process will save time, reduce errors, and enable engineers to focus on more challenging tasks.

Enhance Scalability:
- *Objective:* Ensure the developed solution is scalable to handle a large volume of products with varying descriptions.
- *Rationale*: Addressing scalability concerns will support the efficient classification of the growing number of products.

Drive Business Processes:
- *Objective*: Leverage the trained multi-class classification model to improve various business processes, including product categorization, accuracy, production planning, and cost reduction.
- *Rationale*: The automated solution can extend its impact beyond UNSPSC classification, providing value across multiple business functions.

Reduce Manual Errors:
- *Objective*: Minimize errors associated with manual UNSPSC classification.
- *Rationale*: Automation reduces the risk of human error, ensuring consistency and accuracy in product classification.

Streamline Product Development:
- *Objective*: Streamline product development by automating the classification of easy-to-classify parts, allowing engineers to focus on more complex components.
- *Rationale*: By automating routine tasks, the solution aims to enhance overall efficiency in product development.

Cost Reduction:
- *Objective*: Realize cost reduction through increased efficiency and reduced reliance on manual processes.
- *Rationale*: Automation not only improves accuracy but also optimizes resource allocation, leading to potential cost savings.

Business value/impact

Expected project Benefits:

- Result 1: Cost Savings $25k/year
- Result 2: Additional projects from freed time $100k/year

Total Benefits: $125k/year

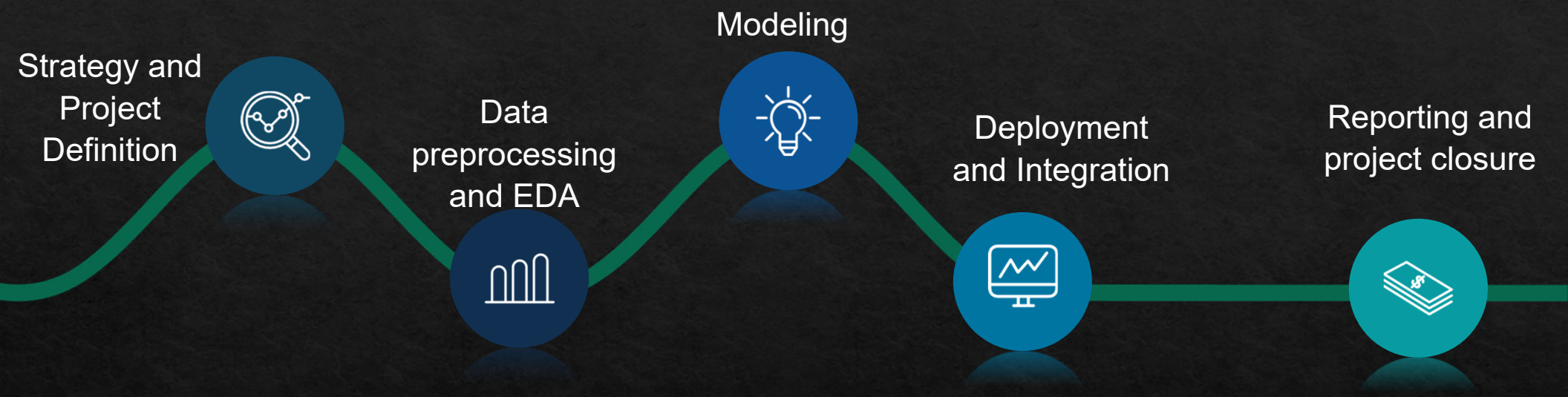
Expected project cost:

1 Year FTE Development cost: $300,000

Expected 3-year return on investment

Project approach and Plan

Ballesteros, Pillai, Li and Tay



Area of Focus	Status	Week 2	Week 3	Week 4	Week 5	Week 6	Week 7	Week 8	Week 9	Week 10
Strategy and project definition	Completed									
Dataset preparation, preprocessing and EDA	Completed									
Modelling	In Process									
Deployment and Integration	Not started									
Reporting and project closure	Not started									


Data

Data overview

- Dataset Size: Approx. 25,000 records.
- Features:
  - Product ID: Unique identifier for each product.
  - Product Description: Textual description of the product.
  - UNSPSC Code: Universal Standard Products and Services Code assigned to each product.
  - Description: UNSPSC description.
- Data sample:

Product #	DESCRIPTIONS	UNPSC CODES	UNPSC Code Description
56828.001	WS-2828,BLK (FAA ONLY)		
PB-PE1250500401-B	L125xW50x0.04mm	811115036	Special packaging
ET-BL0363200001-C	lithium ion battery ;3200 mAH;3.7 V;NTC;Max discharge current=2 C;65x18x- MM;18650 lithium battery , discharge 5A	26111710	Product specific battery packs
ET-BP0375000001-C	lithium polymer battery ;5000 mAH;3.7 V;NTC;Max discharge current=1.5 C;85x55x8.5 MM; Max discharge current=2 C;	26111710	Product specific battery packs
43TX611B0105Z	UPC Label for Box 尺寸: L38.1*W25.4 mm	811115036	Special packaging
1449-63584-162	,0603,5%,1.60K ohm	32121706	Resistor or capacito
1514-26274-002	.ONN,RCPT,STEREO,3.5MM,T/H,GRN		
1457-68643-001	CBL,DB25MM,DBL SHLD,15'	60104912	Electric lead wires or cables
1464-68646-001	SWITCH,OTX STUDIO,MONITOR LIFT	39122221	Switch part or acces
1496-43857-046	IC,SNSR/CNTRL,CAPACITIVE TOUCH		

Data Cleaning

The EDA data cleaning process focuses on removing missing values and duplicated product records in the dataset. Assumptions we make during EDA are:

- Scope of Representation: the dataset may not cover the entire framework of UNSPSC categories, we assume that it comprehensively represents all types of products that HP handles. Each product category within HP's operations is assumed to be present in the data.
- Label Accuracy: The UNSPSC codes associated with each product description in our dataset are accurate.

Data overview

Table 1 Statistics of the dataset

Number of Products	21476
Segments (1st level)	26
Family (2nd level)	52
Class (3rd level)	100
Commodity (4th level)	197
Business Function (5th level)	1
Average ± std of description length	31.1±12.2

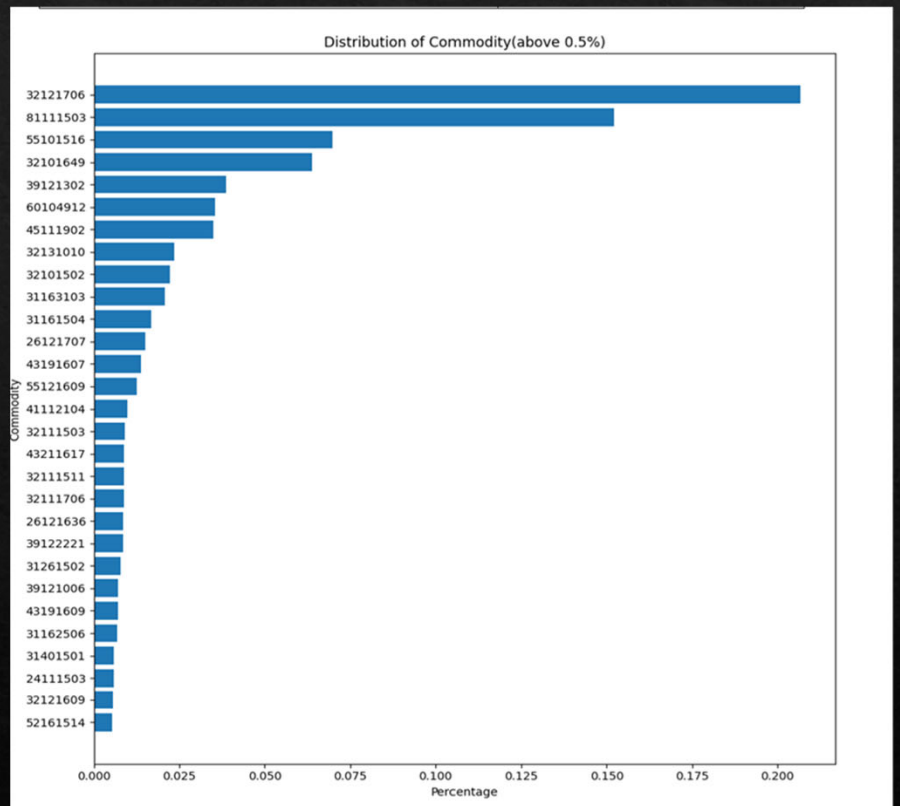


Figure 1 The commodity distribution in the dataset

Feature engineering

NLP techniques that are employed to work with the short text entries：

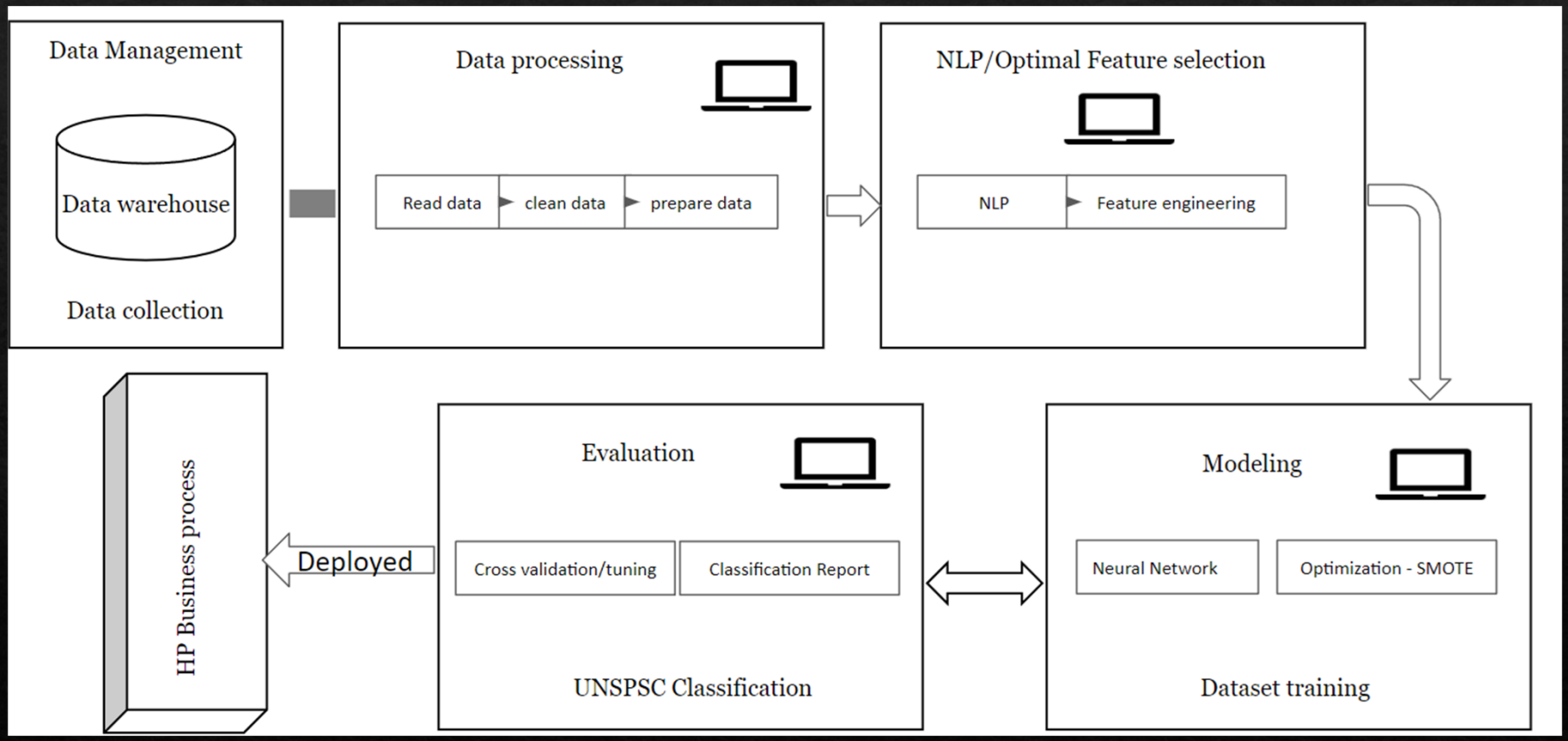
- **Tokenization:** Breaking down a text into individual units. Standalone numbers and those with fractions or percentages were excluded as our customization tokenization process.
- **Text Vectorization:** Convert product descriptions into numerical vectors
  - **Bag-of-Words (BoW)**
  - **Word Embeddings (Word2Vec)**
  - TF-IDF (Term Frequency-Inverse Document Frequency)
  - GloVe embeddings(Global Vectors for Word Representation)

Feature engineering

	Product #	DESCRIPTIONS	UNSPSC CODES	UNSPSC Code Description	Length of DESCRIPTIONS	TOKENS	BAG OF WORDS
0	0	set up sheet for ccx600 d	55101516	Operation or instruction manuals	25	[set, sheet, ccx600]	{'set': 1, 'sheet': 1, 'ccx600': 1}
1	1000.004	res, 270 ohm, cf, 1/8 w, 5%	32121706	Resistor or capacito	27	[res, ohm, cf, w]	{'res': 1, 'ohm': 1, 'cf': 1, 'w': 1}
2	1000.005	res, 200 ohm, cf, 1/8 w, 5%	32121706	Resistor or capacito	27	[res, ohm, cf, w]	{'res': 1, 'ohm': 1, 'cf': 1, 'w': 1}
3	1000.06	res, 56k, cf, 1/8 w, 5%	32121706	Resistor or capacito	23	[res, 56k, cf, w]	{'res': 1, '56k': 1, 'cf': 1, 'w': 1}
4	1000.075	res, 100k, cf, 1/8 w, 5%	32121706	Resistor or capacito	24	[res, 100k, cf, w]	{'res': 1, '100k': 1, 'cf': 1, 'w': 1}
5	1010.005	res, 47 ohm, cf, 1/4 w, 5%	32121706	Resistor or capacito	27	[res, ohm, cf, w]	{'res': 1, 'ohm': 1, 'cf': 1, 'w': 1}
6	1010.01	res, 10 ohm, cf, 1/4 w, 5%	32121706	Resistor or capacito	26	[res, ohm, cf, w]	{'res': 1, 'ohm': 1, 'cf': 1, 'w': 1}
7	1010.025	res, 22 ohm, cf, 1/4 w, 5%	32121706	Resistor or capacito	26	[res, ohm, cf, w]	{'res': 1, 'ohm': 1, 'cf': 1, 'w': 1}
8	1010.043	res, 56 ohm, cf, 1/4 w, 5%	32121706	Resistor or capacito	26	[res, ohm, cf, w]	{'res': 1, 'ohm': 1, 'cf': 1, 'w': 1}

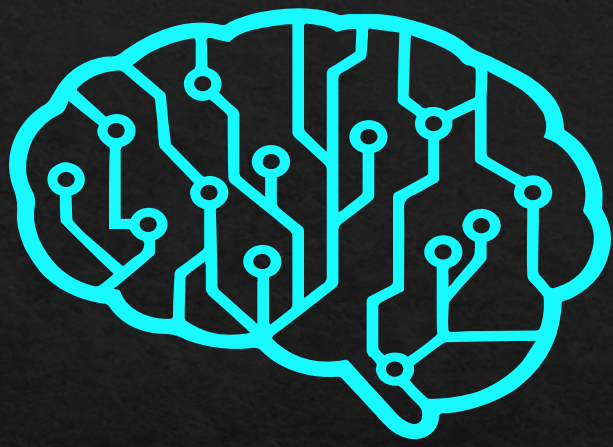
Data flow architecture

Ballesteros, Pillai, Li and Tay




Model

Model Development



Classification algorithms used for each vectorization:

- Logistic Regression
- Multi-Layer Perceptron (MLP) Classifier
- Neural Networks - Keras Sequential API
- Random Forest
- Multinomial Naive Bayes
- Support Vector Machine (SVM)
- XGBoost

Model experiments for each algorithm performed in batches based on vectorization

Model Development - Initial Experiment Setup

Initial evaluation - data distribution and oversampling

- Small scale: selected 50 of the most frequently used UNSPSC codes
- Represents about 92% of the data (19720 out of 21475 usable rows)
- Product # and Description data are combined and oversampled via SMOTE for model training (except for models under GloVe vectorization)
- Initial models utilized default settings for each algorithm

Model Development - Initial Experiment Evaluations

- Models developed under TF-IDF vectorization produced the highest yields in accuracy matching Product # and Description to UNSPSC code for most algorithms
- Next step: TF-IDF vectorization models with hyperparameter tuning
- Top-3 most accurate models after hyperparameter tuning:
  - Logistic Regression (C=100): 95.2%
  - Single-Layer, 128-node, Keras Sequential Neural Network: 95.0%
  - Single-Layer, 2048-node, Keras Sequential Neural Network: 95.1%

		Batch 8							
Experiment		Baseline	2	3	5	6	7	8	9
Predictor	Product #		X	X	X	X	X	X	X
	Product Descriptions	X	X	X	X	X	X	X	X
Data Prep	Tokenization: Using delimiters, Remove stopwords(default)	X							
	Tokenization: Remove digits, fractions, or percentages								
	Tokenization: Vectorization Defaults	X	X	X	X	X	X	X	X
	Vectorization: Bag of Words	X							
	Vectorization: Word2Vec								
	Vectorization: GloVe - glove.6B.100d.txt								
	Vectorization: TF-IDF		X	X	X	X	X	X	X
Oversampling	With Oversampling (SMOTE)		X	X	X	X	X	X	X
	Without Oversampling								
Algorithms	Logistic Regression	X	X						
	Random Forest			X					
	K-Nearest Neighbors								
	MLPClassifier				X				
	Neural Networks					X			
	Multinomial Naive Bayes						X		
	Support Vector Classifier (SVC)							X	
	XGBClassifier								X
Results (Weighted Avg)	Accuracy	0.916	0.9457	0.9447	0.9417	0.9473	0.8829	0.9310	0.9407
	Precision		0.95	0.95	0.94	0.95	0.9	0.94	0.94
	Recall		0.95	0.94	0.94	0.95	0.88	0.93	0.94
	F1- Score		0.95	0.94	0.94	0.95	0.89	0.93	0.94
	Best Accuracy After Parameter Hypertuning (128-node Neural Network)		**0.9518**	0.9462	0.9412	**0.9498**			0.9452
	(2048-node Neural Network)					**0.9508**			

Final Model Evaluation - Experiment Setup

Experiment Limitations utilizing larger data subset

- Many UNSPSC codes in data set are rarely used
- Creating a model trained on all UNSPSC codes (197) is not feasible with the current dataset size
- 114 codes do not provide enough data for model training (707 out of 21475 usable rows, over 3% of the dataset)
- Limit final dataset to UNSPSC codes with at least 20 samples from the usable dataset
- Results in 83 of the most-used UNSPSC codes representing almost 97% of the data (20768 out of 21475 usable rows)
- Provides diverse samples of test and validation sets
- Allows oversampling of training data via SMOTE

Final Model Evaluation - Results

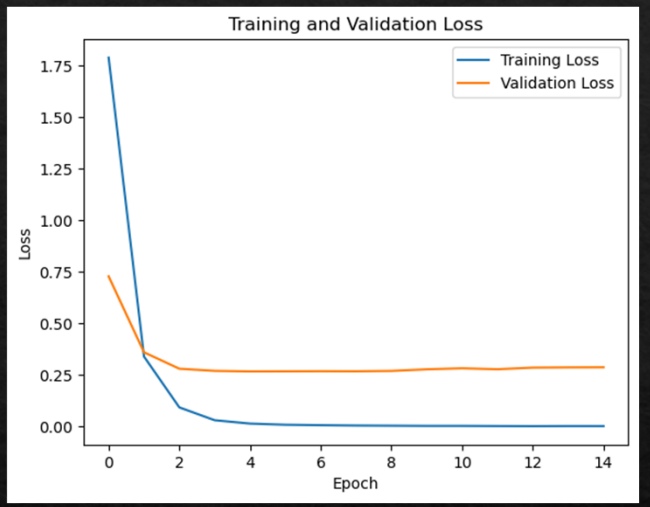
Final evaluation - Results on expanded data subset

- Logistic Regression (C=100): 94.8%
- Single-Layer, 128-node, Keras Sequential Neural Network: 93.6%
- Single-Layer, 2048-node, Keras Sequential Neural Network: 93.5%
- We expected these results to be slightly lower compared to the initial experiments

Final Model Selection

Final evaluation - Considerations for Final Selection

- Expected Logistic Regression to yield higher accuracy due to larger training set and no validation set
- Logistic Regression with C=100: utilizes little to no regularization, highly prone to the model overfitting the training data, and possible poor generalization
- 2048-node neural network: validation loss ceases to improve only after 5 epochs during model fitting, signs of overfitting, possible poor generalization
  - Loss chart at right does not show gradual model fit



Due to overfitting and poor generalization issues in the above models, 128-node Keras Sequential Neural Network is selected as the best overall model


Conclusion

Conclusion

Key Takeaways:

- Successfully developed and deployed a machine learning model to automate UNSPSC code classification.
- Accomplished 93.6% accuracy/performance, exceeding the project's target.
- A significant impact is generated for HP, including a reduction in costs and an increase in operational efficiency.

Next Steps:

- Continue to monitor and optimize the machine learning model.
- Explore opportunities to apply machine learning to other areas of the business.
- Share the project learnings and best practices with the broader community.

W K D QNK
\ R XJ

Sasquatch Engineering team

- Reed Ballesteros
- Manojkumar Damodaran Pillai
- Shawn Tay
- Nan Li