Final Report

AUSTRALIAN MEAT PROCESSOR CORPORATION

Machine Learning for Retention

Machine Learning as a tool to improve retention (Stage 1)

Project Code
2022 - 1138

Prepared by
Katie Scholten

Published by
AMPC

Date Submitted
21/12/2022

Date Published
21/12/2022

Contents

1.0 Executive Summary

A study by the Australian Meat Processor Corporation (AMPC) found staff turnover averaged 63% across the industry and was a significant contributor to a cost competitiveness imbalance which sees Australian labour costs 163% higher than the USA, 238% higher than Argentina and 278% higher than Brazil. Therefore, retaining experienced employees continues to be one of the greatest ongoing challenges within the Australian meat processing sector directly related to higher costs and lost time as well as indirectly to higher incident accident rates.

In this project, Toustone sought to explore the possibility of analysing human resources (HR) data to predict employee retention risk with the expectation of generating insights that will enable active intervention and thus raising staff retention rates and significantly reducing associated costs to industry.

The focus was on the available HR data from one selected red meat processor who agreed to supply data and work with us on developing a model. Further data metrics can be introduced into the model at later date to enhance and improve the model's accuracy.

By analysing, preparing data and building a Machine Learning (ML) model that learns from behaviours of past employees, with the intention to identify like patterns in current employees. The ML model will be instrumental in providing early insights to HR and team managers that allow a proactive conversation to occur and any issues to be addressed well in advance of any unplanned absenteeism or departure.

Two project objectives were agreed upon. These included:

1. To understand if Machine Learning can assist in predicting the likelihood of an employee leaving and hence be a tool to reduce turnover
2. Can Machine Learning identify venerable employees before they leave an organisations?

A further objective was added through a project variation after the success of the first two objectives. This was:

3. To apply the model to two further test sites and thus prove the model’s viability and transferability between processors.

The project methodology was divided into 4 stages:

1. Initial Analysis	2. Data Preparation
◆ Stakeholder engagement ◆ Data acquisition ◆ Discovery	◆ Extract and format data for ML preparation ◆ Undertake data quality checks ◆ Prepare and cleanse data
3. Machine Learning Model Development	**Implement Model at Test Sites**
◆ Design & build suitable cloud infrastructure to support ML ◆ Define hypothesis	◆ Assess transferability to other sites ◆ Produce dashboards and reports ◆ Present outcomes of ML model to each test site

◆ Undertake ML experimentation ◆ Determine most appropriate ML model ◆ Define Precision, Recall & F1 score for selected model ◆ Visualise ML outcomes through BI dashboards	

The scope was defined as exploring retention triggers by employees at two Australian based red meat processing plants owned by a prominent red meat processor. The dataset supplied was for the period 2019 to 2021 and separated into 4 processing dates from which 'as of' analysis could be modelled and included 12 months of data prior to and 12 months of data after each processing date. Data metrics identified and provided were:

- Employee ID
- Gender
- Year of birth
- Employment type
- FTE
- Home postcode
- Leave hours taken
- Employment start date
- Employment
- Employee location
- Employee site
- Visa-Citizenship status
- English as a second language

A binary experiment question was agreed as likely to produce the strongest results given the available data.

Experiment question:

Predict the likelihood that an employee leaves within 12 months

Possible outcomes:

Employee will not be employed after 12 months

Employee will be employed after 12 months

Six algorithm models were explored;

- Logistical regression;
- Random Forest;
- KNN;
- Decision Tree classifier;
- CatBoost; and
- XGBoost.

XGBoost was determined to be the most accurate. With strong results found for the Primary Processor stakeholder, a further 2 processors were invited, and datasets provided. Datasets between each processing site varied slightly and the model was adjusted accordingly.

The base Machine Learning model developed for the Primary Processor was applied to each new processor and an initial fit determined. A quality analysis of each dataset was undertaken before commencing Machine Learning. Some adjustments to the Machine Learning model were applied for each processor to ensure best fit for that organization and dataset. The best fit for the Primary Processor was a XGBoost model while for the 2 Secondary Test Processors, a CatBoost model produced the best results.

The Machine Learning models found:

	Primary Processor	Secondary Test Processor 1	Secondary Test Processor 2
% of the employees who left within 12 months that the model predicted (Recall)	59%	70%	87%
% of which the model accurately predicted those identified as at risk in the recall (Precision)	84%	80%	83%
Overall F1 score (fit of the model, 0 – poor to 1 - perfect)	0.69	0.75	0.85

As a part of best practice requirements, a reporting dashboard in Yellowfin was created to provide a visualization of the data and its relationship to the outcomes discovered in the Machine Learning model. The dashboard included reports examining current and predicted headcount, departure prediction analysis, turnover analysis, leave analysis, departures vs new arrivals analysis and comparative analysis as well as detailed reports highlighting the characteristics of identified employees at risk of leaving within 12 months of analysis.

With the completion of the model customization and the dashboards, an opportunity to explore the Machine Learning outcomes in a way which would provide greater insight into the individuals identified as 'at risk' presented itself.

The initial feature analysis undertaken at the time of the model development showed which metrics contributed the most to the algorithm for each organisation. This highlighted in order of importance which metrics added the most weight to determining 'at risk' individuals. Using this as a base, 7-8 metrics were selected to undertake further analysis. Each metric was categorised as either continuous or categorical and analysed accordingly.

Continuous variables, tenure, age, and base rate of pay, were found to fit a curve which saw probability of leaving reduce as the variable value increased. Analysis of leave taken in the previous 12 months was also analysed for Secondary Test Processor 1. For each organisation and each continuous variable, a fitting function was produced enabling the calculation for the probability for the variable at the individual employee level, highlighting how important that variable is to the determining of an employee being identified as leaving or not leaving within the next 12 months.

Categorical variables, cost centre, employment type, employment site, and contract houses per week, were analysed in turn and a risk rating from Highest to Lowest applied to each category of each variable.

The outcomes of the Machine Learning and dashboards were presented to each Red Meat Processor in turn. Discussions are continuing with one processor in particular keen to explore the viability of operationalizing the model. Another processor is interested in the concept but would like to see a longer-term application of the model through a

12-to-18-month implementation trial before committing to an ongoing operationalization. AMPC support for an implementation trial would be sought should the processor wish to continue to the next stage. Both organisations are keen to broaden the dataset metrics to further enhance the model outcomes.

With AMPC assistance, the next step is to establish a group of 5 plants of various size and structure to deploy the Machine Learning model to. This will not only support the processors in addressing retention issues but also enable ongoing research and refinement of the model. Any Human Resource metrics acquired will also be incorporated, with consent, into AMPC's benchmarking model for the continuing benefit of the industry.

Successful deployment of the model would not only assist the processing sector but also have carry over benefit throughout the rest of the red meat supply chain. If adopted widely within the sector, outcomes could further support producer returns and make the Australian red meat industry more competitive globally.

Based upon these outcomes we conclude that the Machine Learning model as a tool for reducing turnover in red meat processing plants is viable and transferable with minor adjustments for best fit across meat processor plants and organisations.

2.0 Introduction

Background

A study by the Australian Meat Processor Corporation (AMPC) stated that labour-related costs in Australia make up over 58% of the total operating expenses in the meat processing industry. This roughly means that Australia has a 163% higher labour cost than the USA, 238% higher than Argentina and 278% higher than Brazil. Staff industry churn here in Australia is one significant contributor to this cost competitiveness imbalance. Therefore, retaining experienced employees continues to be one of the greatest ongoing challenges within the Australian meat processing sector directly related to higher costs, lost time, and indirectly higher incident accident rates.

In this project, Toustone sought to explore the possibility of analysing HR data to predict employee retention risk with the expectation of generating insights that will enable active intervention thus raising staff retention rates and significantly reducing associated costs to industry.

Given the uncertainties inherent in building a predictive model, this project has been treated as a research and development project supported by AMPC. If successful, the model would not only assist the processing sector but also have carry over benefit throughout the rest of the red meat supply chain. If adopted widely within the sector, outcomes could further support producer returns and make the Australian red meat industry more competitive globally.

Our belief at the commencement of the project was that ML will be a viable option for undertaking analysis of turnover trends in the sector however, as the project has an R&D focus, we were prepared that the direction may change as we understood the data and technology required to deliver the outcome.

The focus was on the available human resources (HR) data from one selected red meat processor who agreed to supply data and work with us on the development of a model. Further data metrics can be introduced into the model at later date to enhance and improve the model's accuracy.

By analysing, preparing data and building a Machine Learning (ML) model that learns from behaviours of past employees, with the intention to identify like patterns in current employees. The ML model will be instrumental in providing early insights to HR and team managers that allow a proactive conversation to occur and any issues to be addressed well in advance of any unplanned absenteeism or departure.

Purpose

The intention of the project is to explore the potential of Machine Learning and its application to Human Resource data to better understand and predict the triggers that adversely affect retention within blue collar work environments. The Machine Learning model is to be developed in such a manner that is repeatable and can be applied to other businesses.

The project will explore the utilization of HR data to predict employee retention risk. It is expected to provide predictive insights that will enable active intervention thus raising staff retention rates and significantly reducing associated costs.

After the development of the Machine Learning model, a variation to the project was applied for and approved. In addition to developing the model, the variation also seeks to apply it to two further processing plants to ascertain the viability of the model to transfer between different operations and meat processors.

3.0 Project Objectives

Two project objectives were agreed upon. These included:

1. To understand if Machine Learning can assist in predicting the likelihood of an employee leaving and hence be a tool to reduce turnover
2. Can Machine Learning identify venerable employees before they leave an organisations?

A further objective was added through a project variation after the success of the first two objectives. This was:

3. To apply the model to two further test sites and thus prove the model's viability and transferability between processors.

4.0 Methodology

The project methodology was broken into 4 distinct phases.

Initial Analysis

- Stakeholder engagement
- Data acquisition
- Discovery

Data Preparation

- Extract and format data for Machine Learning preparation
- Undertake data quality checks
- Prepare and cleanse data

Machine Learning Model Development

- Design and build cloud infrastructure to support Machine Learning development, training, and implementation
- Define hypothesis
- Undertake Machine Learning experimentation including training and testing of models
- Determine most appropriate Machine Learning model
- Define Precision, Recall, and F1 score for the selected model
- Produce Dashboards and Reports to visualise Machine Learning outcomes

Implement Model at Test Sites

- Assess the transferability of the Machine Learning model by applying to datasets from two additional Meat Processors
- Produce Dashboards and Reports to visualise Machine Learning outcomes for each test Meat Processor
- Present outcomes of Machine Learning model to each test Meat Processor

5.0 Project Outcomes

The project commenced with the engagement of two stakeholders. A data science consulting firm to undertake the Machine Learning model development and a prominent red meat processor to provide data and functional expertise.

The scope was defined as exploring retention triggers by employees at two Australian based red meat processing plants owned by the processor stakeholder. Data metrics identified and provided were:

- Employee ID
- Gender
- Year of birth
- Employment type
- FTE
- Leave hours taken
- Employment start date
- Employee site
- Employee location
- Current role (lack of data consistency precluded its inclusion in the ML model)
- English as a second language
- Visa-Citizenship status
- Home postcode
- Date of departure

Exit and engagement survey, and injury data was identified as being ‘nice to have’ but difficult to source.

The dataset supplied was for the period 2019 to 2021 and separated into 4 processing dates from which ‘as of’ analysis could be modelled and include 12 months of data prior to and 12 months of data after each processing date.

An experiment question was developed, initially looking at multiple timeframes and outcomes in which an employee was likely to leave their employment but, in the end, the strongest results were produced with a binary hypothesis.

Experiment question:
Predict the likelihood that an employee leaves within 12 months

Possible outcomes:
Employee will not be employed after 12 months
Employee will be employed after 12 months

The experimentation was conducted in three phases.

1. Initial testing

   Initial testing was designed to be quick with minimal overheads and to inform data preparation changes should any need to be made. At this stage initial insight into target variable usage and training times were determined.

2. Multiple model exploration

   With information from initial testing, a range of algorithms was explored over moderately deep training to determine 2-3 models with the most potential. General accuracy metrics were captured for each model and compared.

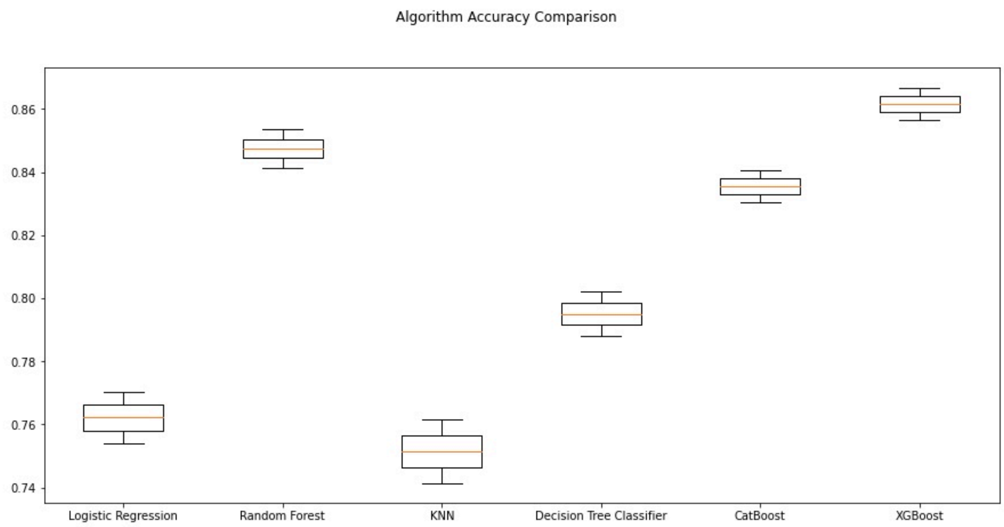
3. Top model deep dives and optimisation

   The top 2-3 models were selected and underwent hyperparameter tuning and deep training by utilising sampling techniques and exploring specific accuracy metrices per model type. This testing aimed to give the best performance for each model within the problem constraints and was used to directly inform the validity of the experiment and predictive qualities of the dataset.

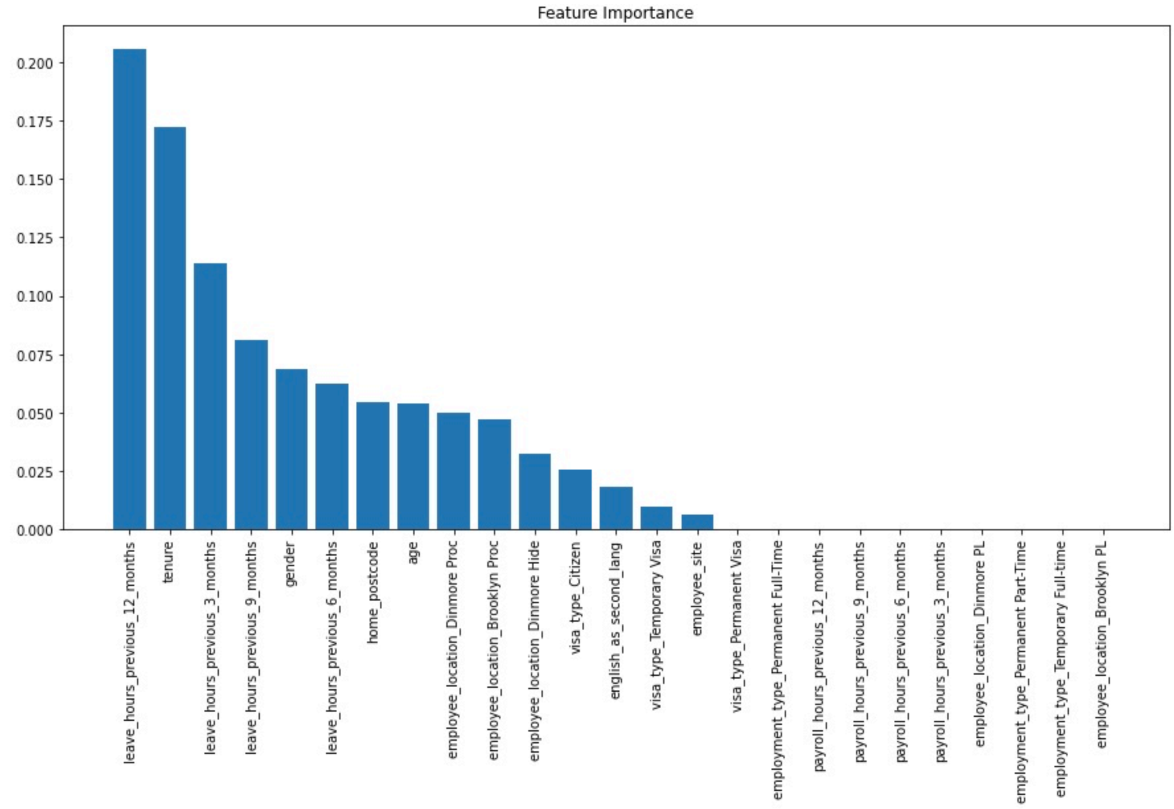
Six algorithm models were explored;

- Logistical regression;
- Random Forest;
- KNN;
- Decision Tree classifier;
- CatBoost; and
- XGBoost.

XGBoost was determined to be the most accurate.



Feature importance was examined determining the metrics with the most statistical weight in determining the outcomes of the algorithm.



An invitation to a further meat processors was issued with two agreeing to supply datasets.

There were slight variations between the datasets supplied by all three meat processors. A comparison across processors is presented below.

Data Metric	Primary Processor	Secondary Test Processor 1	Secondary Test Processor 2
Leave hours taken	Y	Y	Y
Tenure	Y	Y	Y
Age	Y	Y	Y
Visa/Citizenship status	Y	Y	Y
Employment type	Y	Y	Y
Processing site	Y	Y	Y
Gender	Y	Y	Y
English as a 2nd language	Y		
Home postcode	Y	Y	
Payroll hours worked	Y		Y
Current role		Y	
Work location within plant		Y	Y
Team/Department		Y	Y
Injury count		Y	

Base payrate		Y	Y

The base Machine Learning model developed for the Primary Processor was applied to each Secondary Test Processor and an initial fit determined. A quality analysis of each dataset was undertaken before commencing Machine Learning. Some adjustments to the Machine Learning model were applied for each processor to ensure best fit for that organization and dataset. The best fit for the Primary Processor was a XGBoost model while for both Secondary Test Processors, a CatBoost model produced the best results.

The Machine Learning models found:

	Primary Processor	Secondary Test Processor 1	Secondary Test Processor 2
% of the employees who left within 12 months that the model predicted (Recall)	59%	70%	87%
% of which the model accurately predicted those identified as at risk in the recall (Precision)	84%	80%	83%
Overall F1 score (fit of the model, 0 – poor to 1 - perfect)	0.69	0.75	0.85

Finally visual representation of the data was developed with the production of business intelligence dashboards.

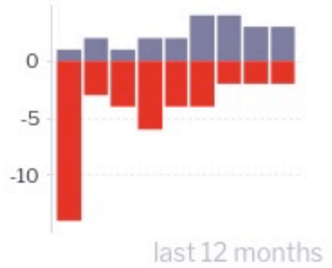
6.0 Discussion

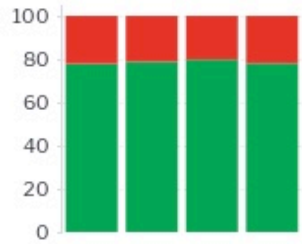
What does this mean for business?

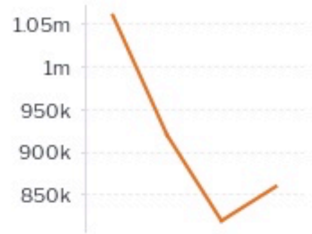
- Identifies employees at risk of leaving (list of employee IDs produced)
- Provides an opportunity to work directly with at risk employees
- Provides insight into what increases the likelihood of employees leaving
- Provides insight into why employees choose to stay
- Enables data driven decisions on changes and initiatives in the business that can be implemented to reduce employee turn over
- Saves money – recruitment, processing, and productivity costs
- The model is transferable to different operations and meat processors with limited customization required provided the core data metrics with sufficient history is provided.

As a part of best practice requirements, a reporting dashboard in Yellowfin was created to provide a visualization of the data and its relationship to the outcomes discovered in the Machine Learning model. The dashboard included reports examining current and predicted headcount, departure prediction analysis, turnover analysis, leave analysis, departures vs new arrivals analysis and comparative analysis as well as detailed reports highlighting the characteristics of identified employees at risk of leaving within 12 months of analysis.

Current Turnover Rate	Current ML Prediction	Cost of Predicted Turnover
20.7%	**78.1%**	**860k**
Avg Headcount	Percentage **Remain**	* Based on estimated **$20k turnover cost** per departure.
198	**21.9%**	Total **Depart** Headcount
Departures 41 / New Hires 22	Percentage **Depart**	43



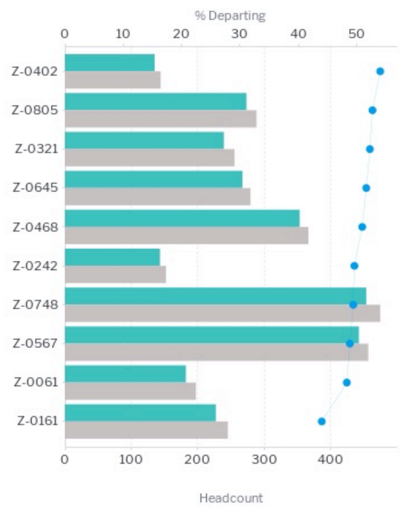




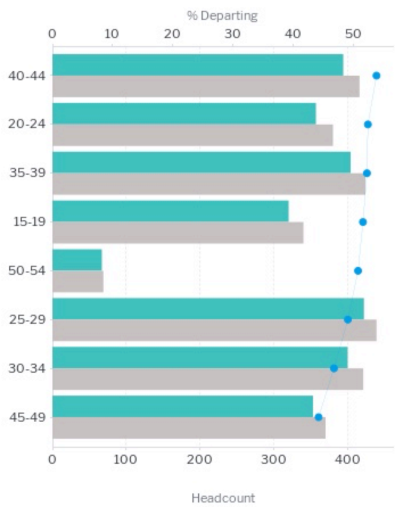
Departure Prediction Analysis

Display the **current** headcount (**top bar**) compared to **last year** headcount (**bottom bar**). The **dot** in each chart displays the **predicted departure percentage**.

By Manager Id

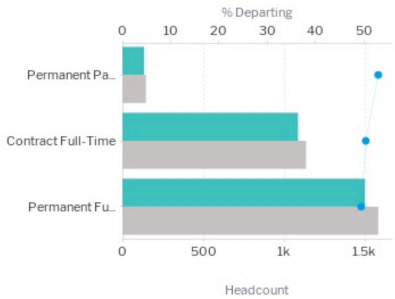


By Age Group

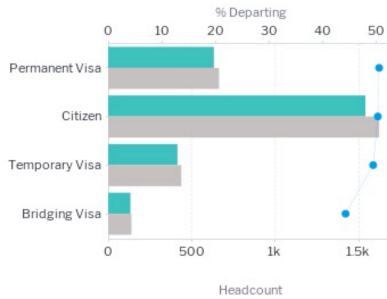


By Role

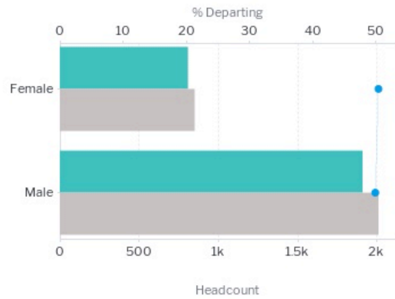
By Employment Type



By Visa Type



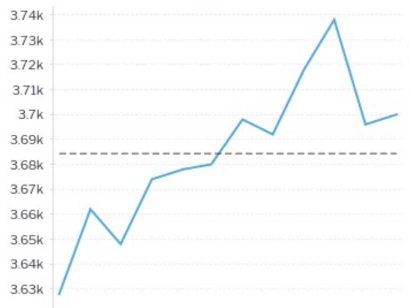
By Gender



Overview Turnover Leave Departures & New Hire Analysis

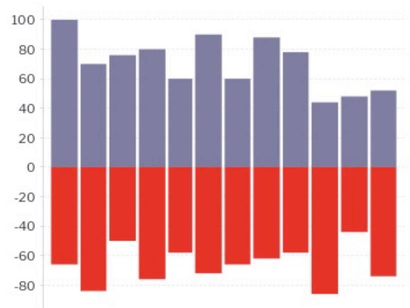
Headcount by Month

Display the **total** headcount **by month** with the **average** trend line.



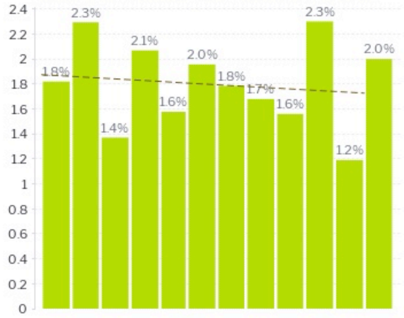
Departures & New Hires by Month

Display the **total** departures and new hires **by month**.



Departure vs Headcount Rate

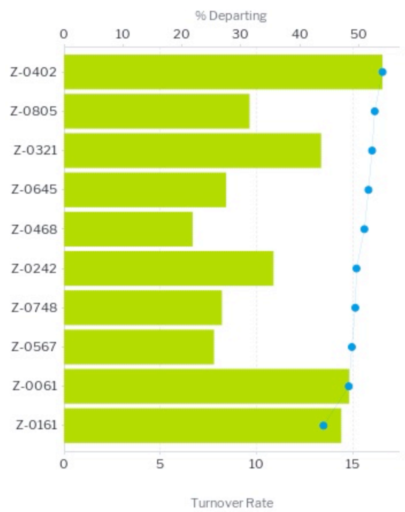
Display the **total** departures as a **percentage** of the headcount **by month** with the **linear** trend line.



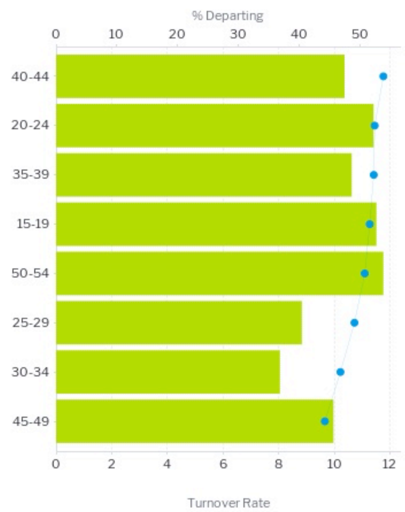
Departure Prediction Analysis

Display the **turnover rate** compared to the **predicted departure percentage**.

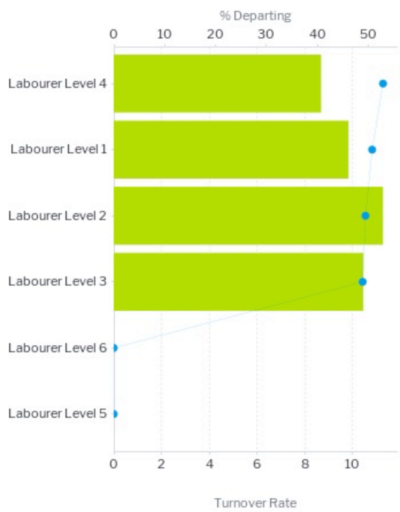
By Manager Id



By Age Group



By Role



Current At Risk Analysis

Display the current list of at risk employees based on specific filters.

1 - 100 / 144

Site	Location	Manager	Team	Team Leader	Current Role	Employee Id	ML Outcome	Employment Type	Gender	Visa Type	Age Group	Leave Hours
Western Plant	W_Skins	Z-0805	WS1	A-0761	Labourer Level 1	I-4760	At Risk	Permanent Full-Time	Male	Bridging Visa	20-24	152.00
						O-0766	At Risk	Permanent Full-Time	Female	Citizen	45-49	106.40
						P-4760	At Risk	Contract Full-Time	Male	Temporary Visa	20-24	0.00
						X-0765	At Risk	Permanent Full-Time	Male	Permanent Visa	20-24	706.80
					Labourer Level 2	A-0768	At Risk	Permanent Full-Time	Male	Citizen	45-49	311.60
						E-0765	At Risk	Permanent Full-Time	Male	Citizen	35-39	608.00
						F-0763	At Risk	Contract Full-Time	Male	Citizen	35-39	790.40
						G-2760	At Risk	Contract Full-Time	Female	Bridging Visa	35-39	45.60
						H-0761	At Risk	Permanent Full-Time	Male	Citizen	25-29	258.40
						H-0767	At Risk	Contract Full-Time	Female	Citizen	15-19	912.00
						I-0766	At Risk	Permanent Full-Time	Male	Permanent Visa	20-24	182.40
						L-4761	At Risk	Permanent Full-Time	Male	Permanent Visa	45-49	190.00
						M-0764	At Risk	Contract Full-Time	Female	Permanent Visa	20-24	440.80
						R-0760	At Risk	Contract Full-Time	Male	Permanent Visa	20-24	243.20

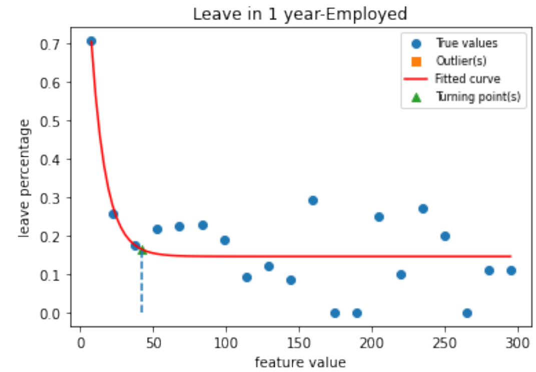
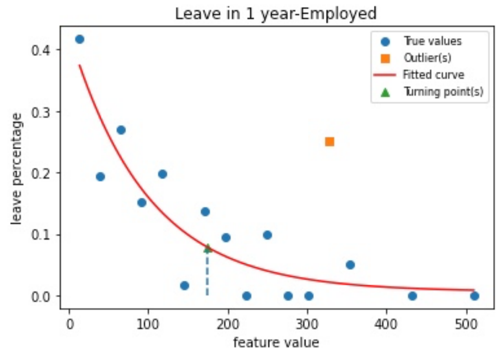
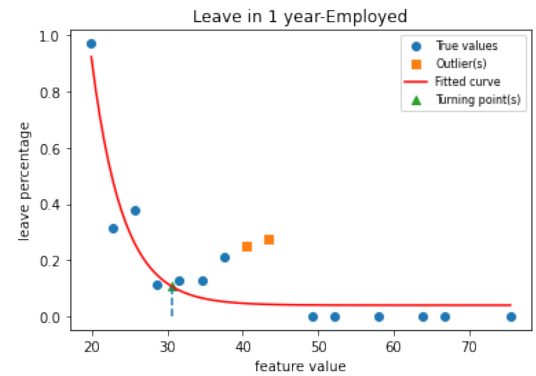
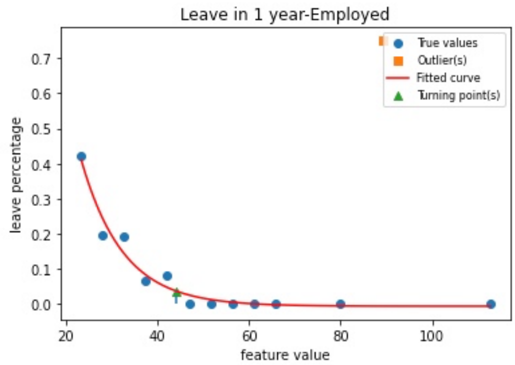
With the completion of the model customization and the dashboards, an opportunity to explore the Machine Learning outcomes in a way which would provide greater insight into the individuals identified as ‘at risk’ presented itself.

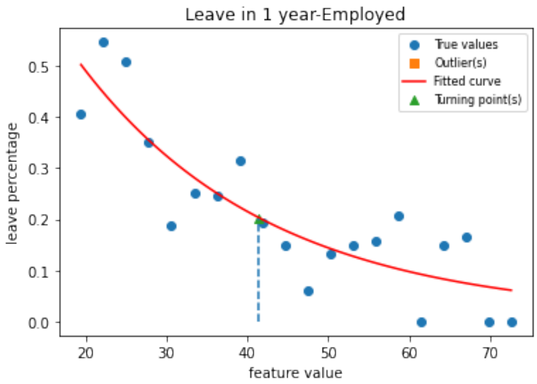
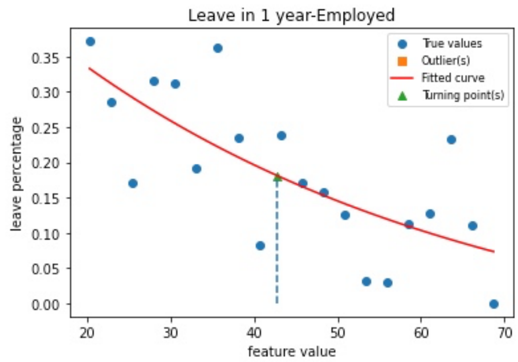
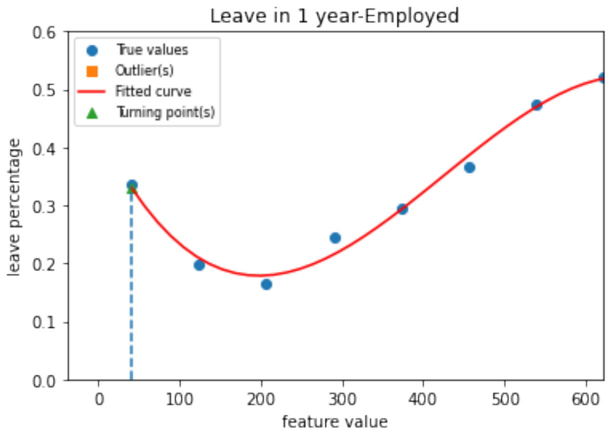
The initial feature analysis undertaken at the time of the model development showed which metrics contributed the most to the algorithm for each organisation. This highlighted in order of importance which metrics added the most weight to determining ‘at risk’ individuals. Using this as a base, 7 metrics were selected to undertake further analysis. Each metric was categorised as either continuous or categorical and analysed accordingly.

Continuous variables; tenure; age; and base rate of pay; were found to fit a curve which saw probability of leaving reduce as the variable value increased. Analysis of leave taken in the previous 12 months was also analysed for Secondary Test Processor 1. For each organisation and each continuous variable, a fitting function was produced enabling the calculation for the probability for the variable at the individual employee level, highlighting how important that variable is to the determining of an employee being identified as leaving or not leaving within the next 12 months.

Outcomes of Feature Analysis – continuous variables:

Secondary Test Processor 1	Secondary Test Processor 2
Tenure	

<table>
<tr><td>
</td><td>
</td></tr>
<tr><td>Turning point for tenure_in_months is 42.5 months (approx. 3.5 years). This indicates that for employees hired less than 3.5 years, their leave probability drops quickly with the increase in tenure. On the other hand, for employees hired more than 3.5 years, their probability will remain at a lower level and not change much.</td><td>Turning point for tenure_in_months is 174 months (approx. 14.5 years). This indicates that for employees hired less than 14.5 years, their leave probability drops quickly with the increase in tenure. On the other hand, for employees hired more than 14.5 years, their probability will remain at a lower level and not change much.</td></tr>
<tr><td colspan="2">Base rate of pay</td></tr>
<tr><td>
</td><td>
</td></tr>
<tr><td>Turning point for StdRate is $30.58 per hour. This indicates that for employees paid less than $30.58 per hour, the probability of employees leaving decreases as the per hour rate increases until they reach $30.58, after which leave probability will remain at a lower level and not change much.</td><td>Turning point for StdRate is $44.17 per hour. This indicates that for employees paid less than $44.17 per hour, the probability of employees leaving decreases as the per hour rate increases until they reach $44.17, after which leave probability will remain at a lower level and not change much.</td></tr>
<tr><td colspan="2">Age</td></tr>
</table>

<table>
<tr>
<td>
</td>
<td>
</td>
</tr>
<tr>
<td>Leaving percentage decreases as age increases. The turning point in this analysis is less meaningful since there is no obvious knee points in the fitting curve.</td>
<td>Leaving percentage decreases as age increases. The turning point in this analysis is less meaningful since there is no obvious knee points in the fitting curve.</td>
</tr>
<tr>
<td colspan="2">Leave hours taken in last 12 months</td>
</tr>
<tr>
<td>
</td>
<td>Not available</td>
</tr>
<tr>
<td>With leave hours increasing, leaving percentage decreases first and then starts to increase from approximately 200 hours in the previous year. The turning point is meaningless in this analysis.</td>
<td></td>
</tr>
</table>

Categorical variables, cost centre, employment type, employment site, and contract houses per week, were analysed in turn and a risk rating from Highest to Lowest applied to each category of each variable.

Example of feature analysis reporting for Secondary Test Processing 2 is presented below.

Employee Risk Detail Report (Drill Through)

Headcount & Departure Prediction Percentage by Age Group (Drill Through) > Employee Risk Detail Report (Drill Through)

Redacted for privacy

Employee Id	Location	Current Role	Age	ML Outcome	Tenure Risk	Age Risk	Std Rate Risk	Cost Account Risk Group	Employee Site Risk Group	Employment Type Risk Group	Contract Week Hrs Risk Group
			24	Staying	20.74	30.29	20.81	Lowest	Lower	Lower	Lower
			21	Staying	24.24	32.70	14.01	Lower	Lower	Lower	Lower
			23	At Risk	25.80	31.08	14.01	Lower	Lower	Lower	Lower
			23	Staying	24.47	31.08	26.04	Lower	Lower	Lower	Lower
			21	At Risk	28.12	32.70	40.66	Lower	Lower	Lower	Lower
			22	Staying	28.87	31.88	32.20	Lower	Lower	Lower	Lower
			21	Staying	29.37	32.70	29.12	Lower	Lower	Lower	Lower
			23	At Risk	32.76	31.08	34.42	Lowest	Lower	Lower	Lower
			20	Staying	30.29	33.53	26.04	Lower	Lower	Lower	Lower
			22	Staying	30.44	31.88	21.11	Lowest	Lower	Lower	Lower
			24	Staying	30.52	30.29	22.01	Lower	Lower	Lower	Lower
			21	Staying	30.73	32.70	32.20	Lower	Lower	Lower	Lower
			20	At Risk	35.25	33.53	38.47	Lower	Lower	Lower	Lower
			20	Staying	31.42	33.53	34.42	Higher	Higher	Lower	Lower

7.0 Conclusions / Recommendations

The outcomes of the Machine Learning and dashboards were presented to each Red Meat Processor in turn.

Based upon these outcomes we conclude that the Machine Learning model as a tool for reducing turnover in red meat processing plants is viable and transferable with minor adjustments for best fit across meat processor plants and organisations.

Discussions are continuing with one processor keen to explore the viability of operationalizing the model. Another processor is interested in the concept but would like to see a longer-term application of the model through a 12-to-18-month implementation trial before committing to an ongoing operationalization. AMPC support for an implementation trial would be sought should the processor wish to continue to the next stage. Both organisations are keen to broaden the dataset metrics to further enhance the model outcomes.

With AMPC assistance, the next step is to establish a group of 5 plants of various size and structure to deploy the Machine Learning model to. This will not only support the processors in addressing retention issues but also enable ongoing research and refinement of the model. Any Human Resource metrics acquired will also be incorporated, with consent, into AMPC's benchmarking model for the continuing benefit of the industry.

Successful deployment of the model would not only assist the processing sector but also have carry over benefit throughout the rest of the red meat supply chain. If adopted widely within the sector, outcomes could further support producer returns and make the Australian red meat industry more competitive globally.

8.0 Bibliography

Not applicable

9.0 Appendices

None