Predicting propensity to buy at Pedal Power

Data analysis, interpretation and prediction

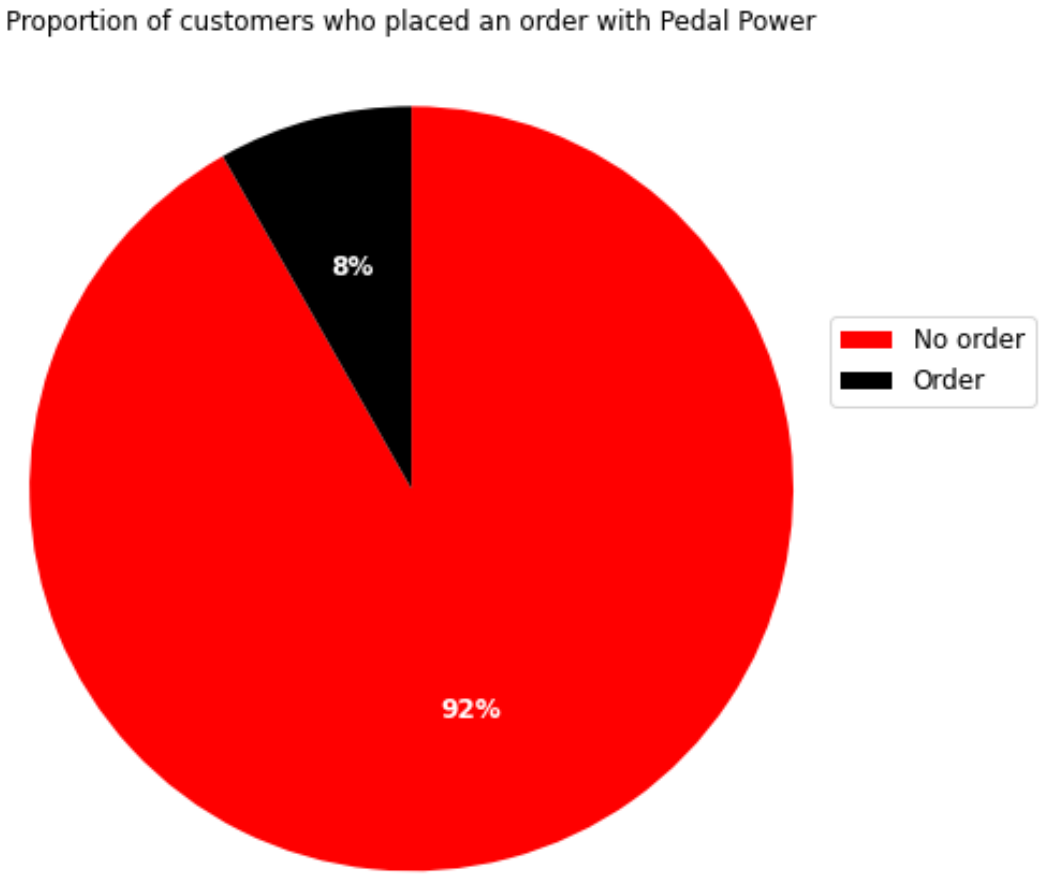
by Amy Birdee

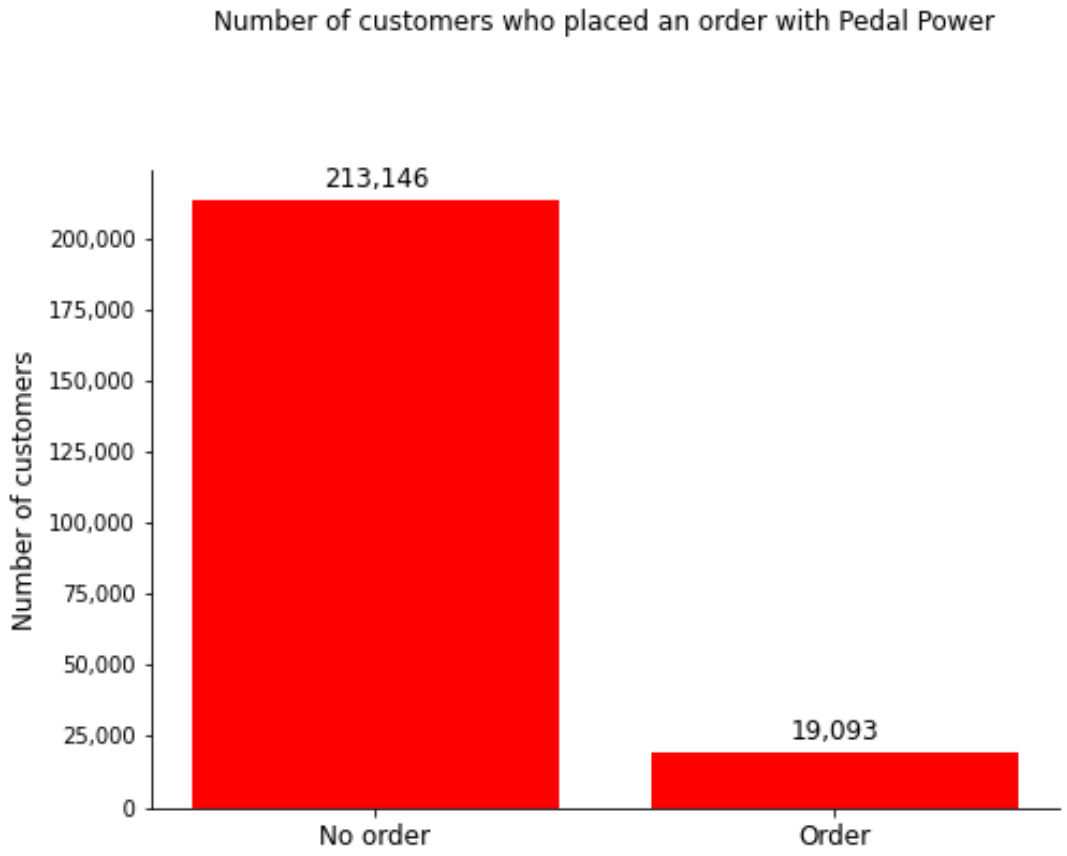
Introduction

- The data consist of details of customers at Pedal Power, a hypothetical company selling home exercise bikes
- There are over 200,000 rows of customer data. I have assumed the data span one week
- Included in the data are variables such as whether or not a customer logged into their account, whether they are a returning user, whether they saw the checkout page and whether or not they placed an order for a bike
- This project aims to segment the customer data and build a classifier model which will predict whether or not a customer will place an order
- The machine learning models used in the project are a logistic regression and Random Forest. The hyper-parameters of the Random Forest model were tuned in an attempt to build an optimal model

How many customers placed an order at Pedal Power?

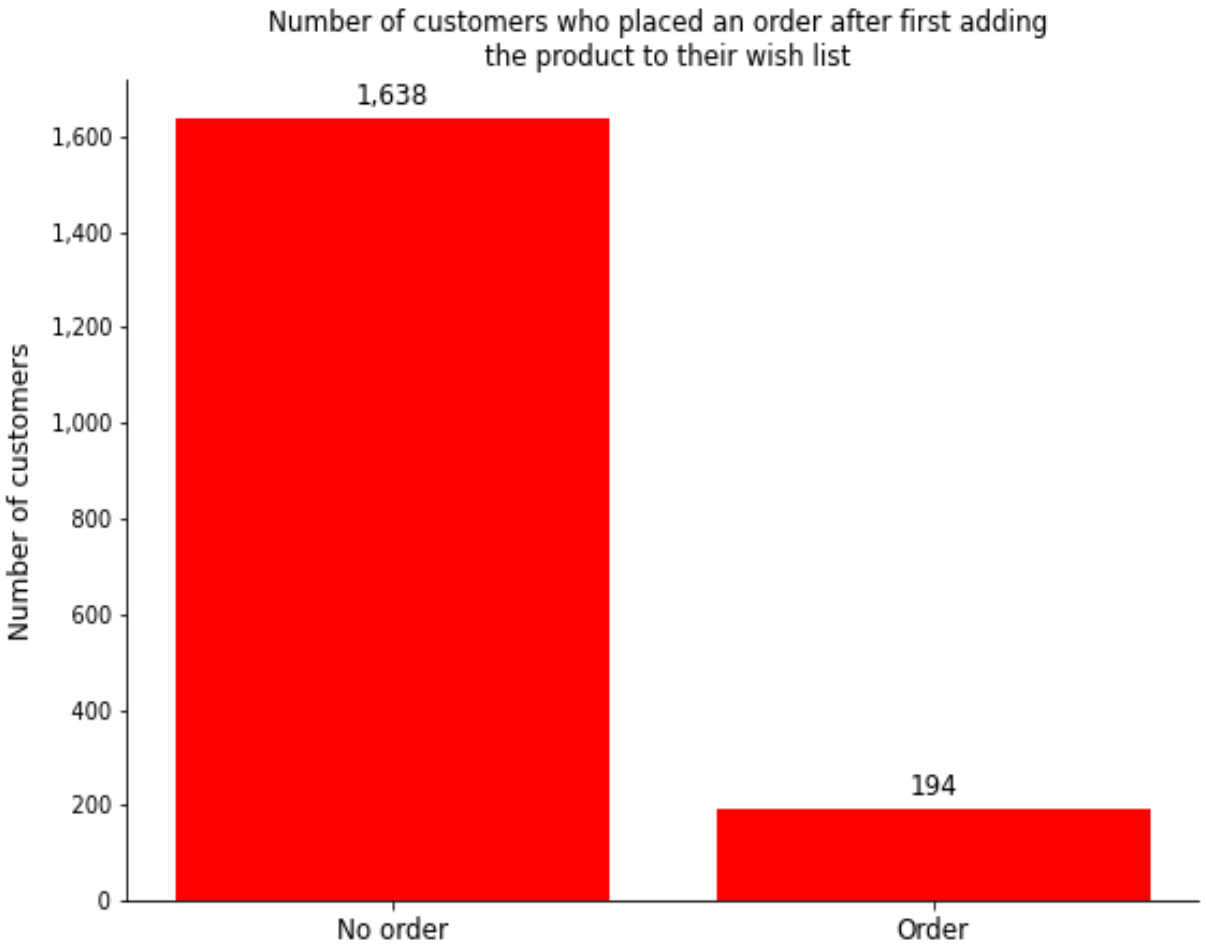
- Only **8%** of customers placed an order in the last week which equates to just over **19,000** customers





- This suggests the exercise bikes are high ticket items and quite expensive - certainly not an impulse buy

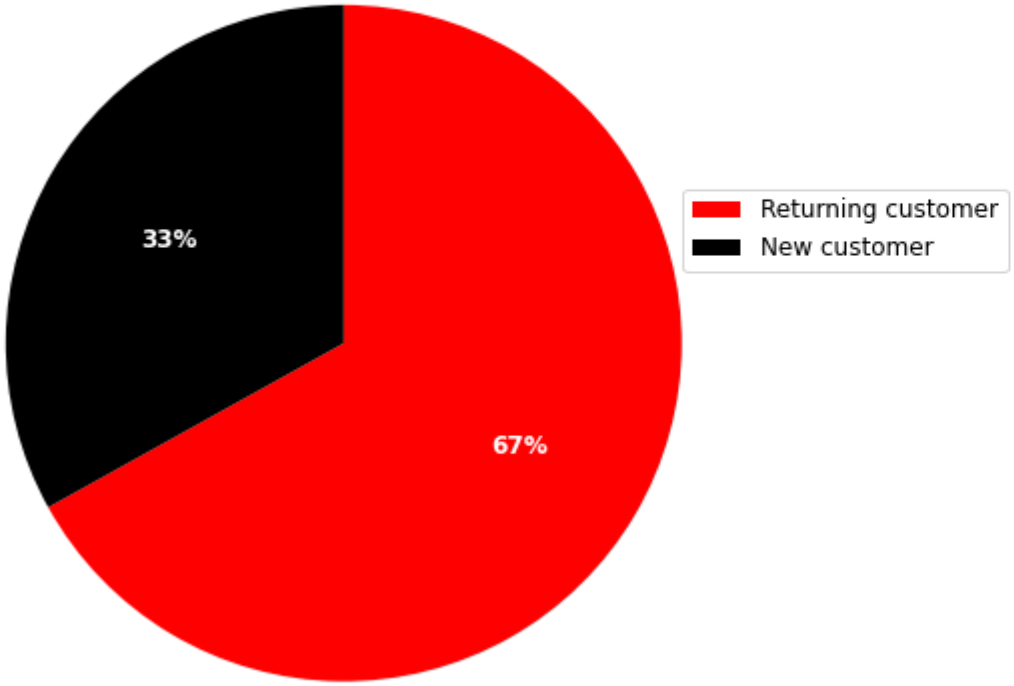
How does adding a product to a wish list affect orders?

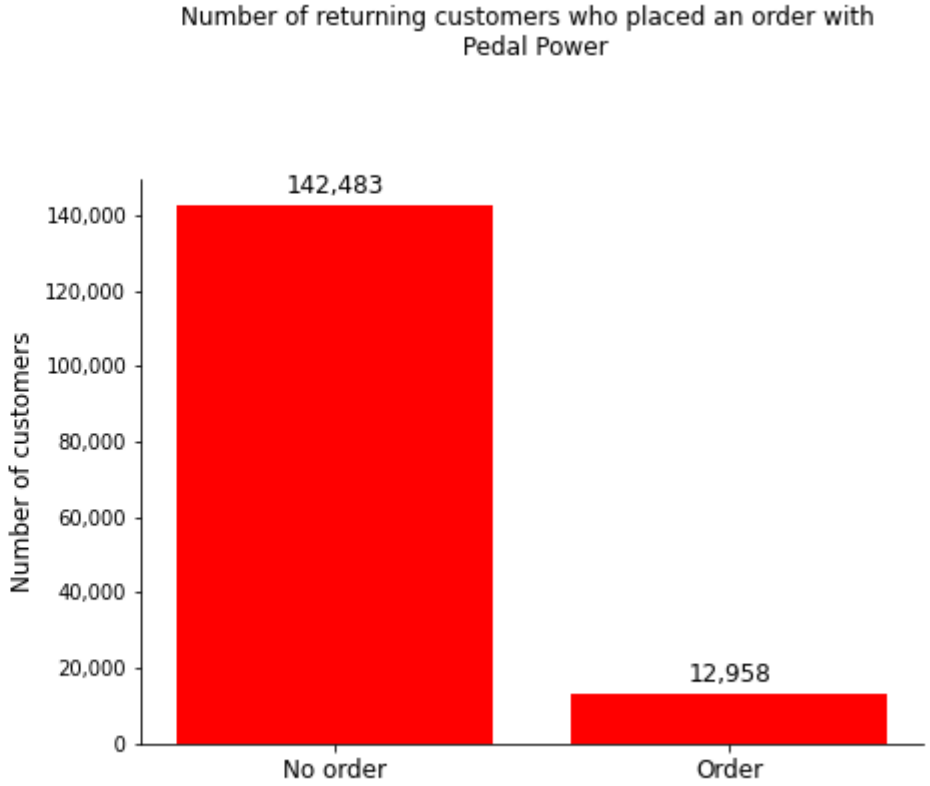


- Some customers may add a bike to their wish list to buy at a later date when they have more time or money or both
- Of all customers who added a product to their wish list, **11%** then went on to place an order
- This is still quite a low percentage and so adding to the wish list may not be indicative of placing an order
- There are **1,638 potential customers** who have added a bike to their wish list but have not yet placed an order – Pedal Power could try **targeting advertising** at these users to encourage them to buy

Do returning users place more orders?

- Pedal Power has a high proportion of returning customers at **67%**. Whilst not all of them place an order, this does indicate **loyalty to the brand**

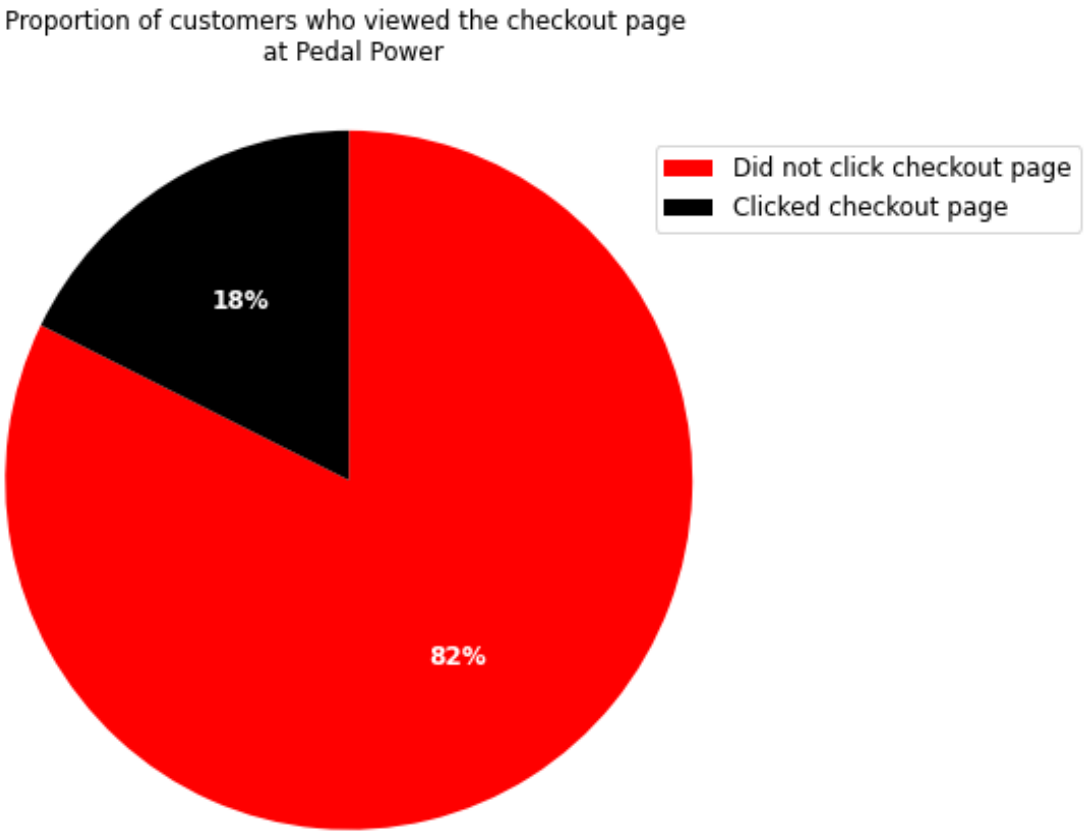




- Of all returning customers, **8%** placed an order in the last week. Given that many customers are returning customers, they **may place orders in the coming weeks** – perhaps they are returning because they haven't made a decision on their ideal bike yet and need to do some further research

Do all customers who reach the checkout page place an order?

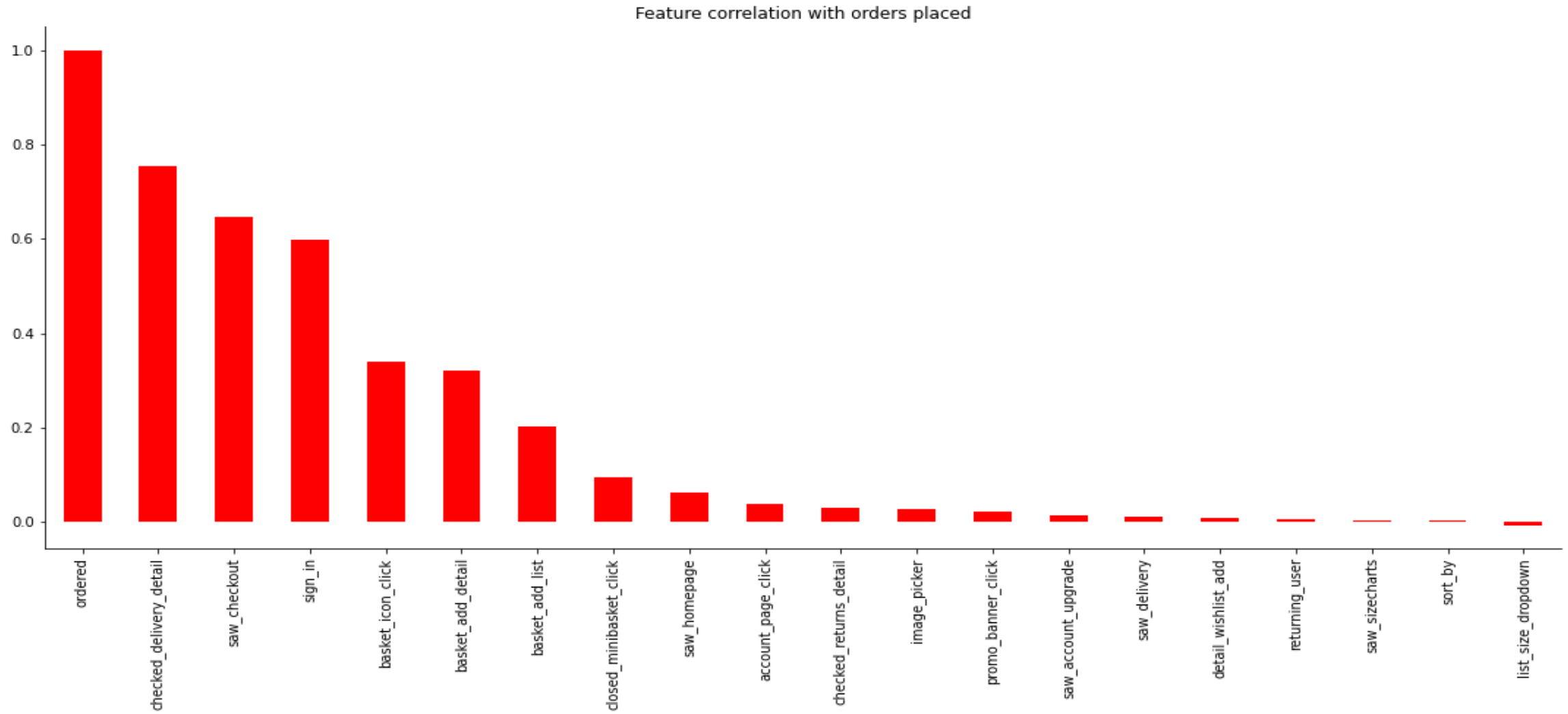
- Only **18%** of the 232,000 website visitors reached the checkout page but not all of these visitors placed an order



- However, a high proportion of those who do visit the checkout page do place an order – **46%** of customers who saw the checkout page placed an order. Therefore, clicking the checkout page is likely indicative of a customer placing an order

How do all features of the dataset correlate with orders?

- Checking the delivery detail and viewing the checkout page are most positively correlated with orders whereas viewing a size dropdown and sorting products are least correlated



Evaluating logistic regression model: confusion matrix

- The aim of the logistic regression analysis is to **reduce the number of false positives**, i.e. those customers **we predict will place an order but they do not**

- The logistic regression model classified 687 instances as false positives out of the 46,448 customers included in the test dataset

		Prediction: No order (0)	Prediction: Order (1)
Actual	No order	True negative (TN)	**False positive (FP)**
Actual	Order	False negative (FN)	True positive (TP)

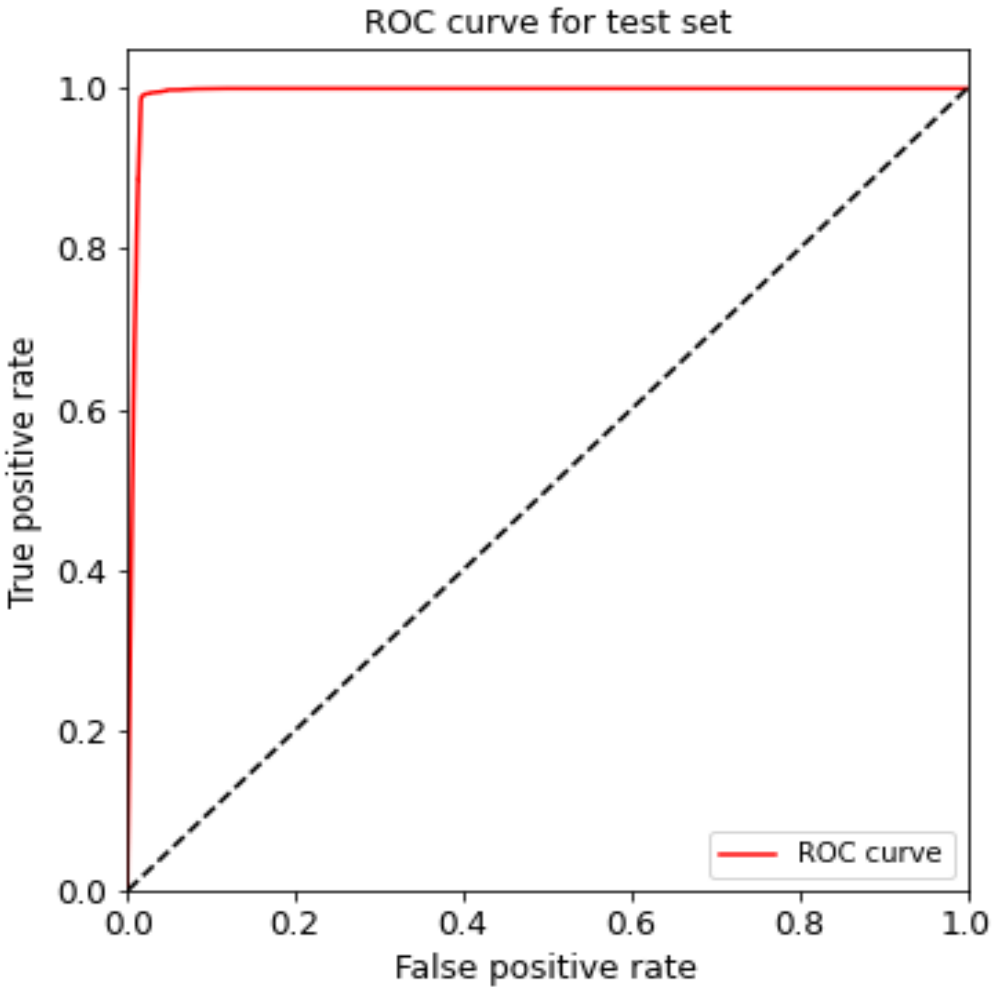
		Prediction: No order (0)	Prediction: Order (1)
Actual	No order	41,957	**687**
Actual	Order	56	3,748

Evaluating logistic regression model: key metrics

- In a propensity to buy analysis, we want to **reduce the instances of false positives**. Therefore, we are trying to optimise the **precision** metric which shows how precise the predictions are
- In other words, **out of the times the model said a customer would place an order, how may times did they actually order?**
- The precision score for the logistic regression model is **85%**

	Precision	Recall	Accuracy	F1 score
Calculation	TP / (TP + FP)	TP / (TP + FN)	(TP + TN) / (TP + FP + TN + FN)	2 * (Precision * Recall) / (Precision + Recall)
Result	**0.85**	0.99	0.98	0.91

Evaluating logistic regression model: ROC curve and AUC



- The ROC (receiver operator characteristic) curve is another way to **evaluate the model** visually
- The true positive rate is mapped against the false positive rate of the classifier.
- The best models will have an **ROC curve that hugs the upper left corner of the graph** – i.e. the model correctly classifies the positives more often than it incorrectly classify them.
- The curve here is in the **upper portion of the grid and also far from the 50% line** which is encouraging and suggests the model is very robust
- The AUC (area under curve) score is **0.99** which is also a great score (AUC score can be between 0 and 1)

Evaluating Random Forest models: confusion matrix

- The data were run through **two Random Forest models** – one was a default model with no hyper-parameter tuning and the other had four hyper-parameters tuned. These were: num_estimators, max_depth, min_samples_split and min_samples_leaf

- The tuned model **performed worse at minimising false positives**. The logistic regression model performed best at limiting false positives

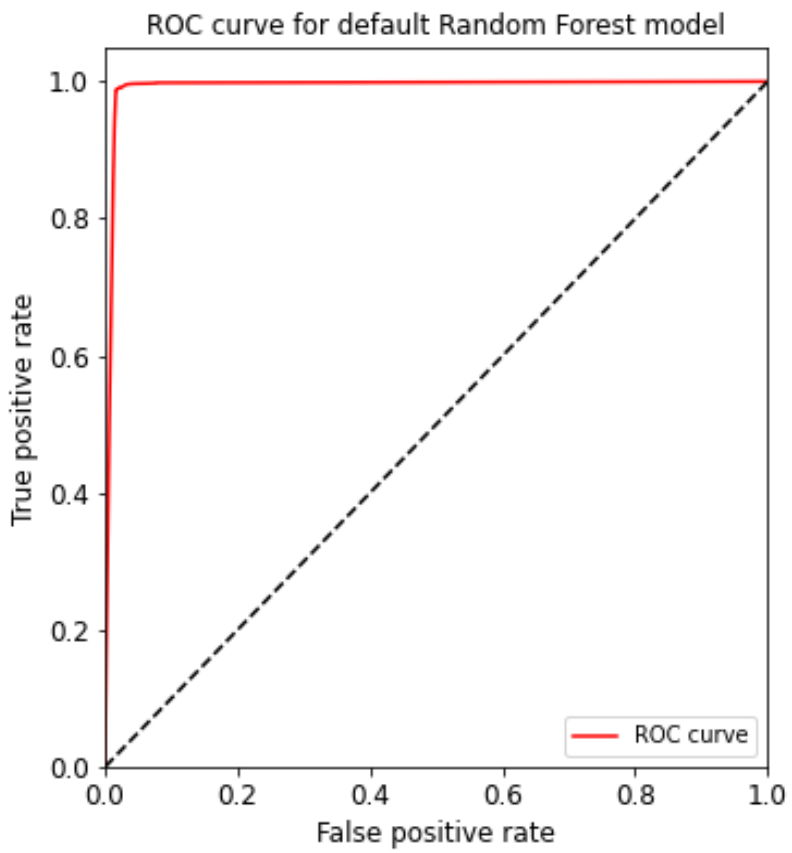
Default model

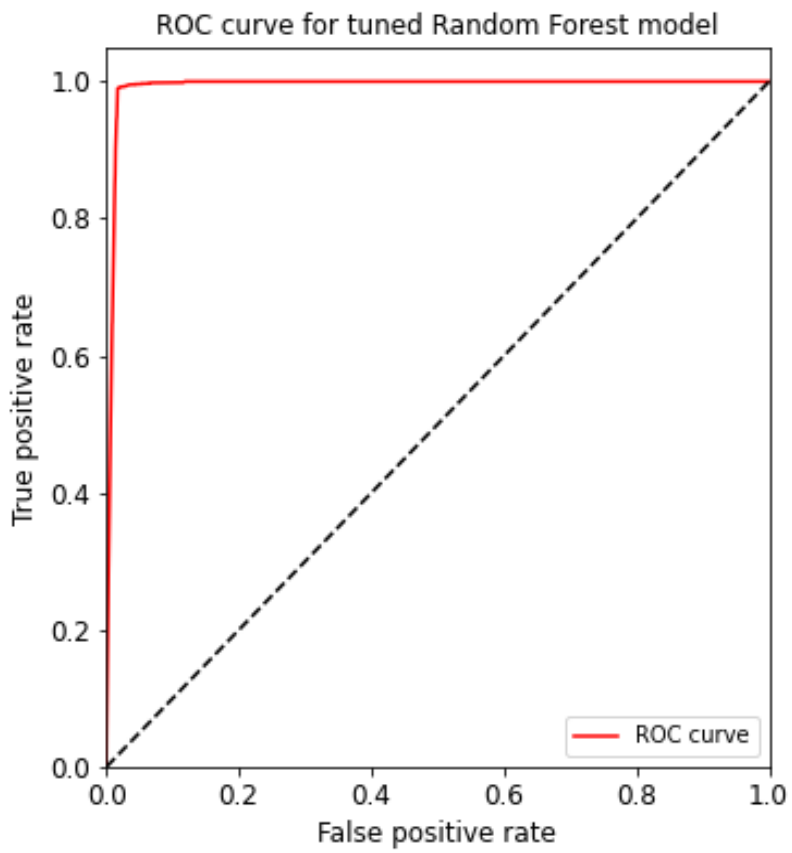
Actual \ Prediction	No order (0)	Order (1)
No order	41,975	**669**
Order	101	3703

Tuned model

Actual \ Prediction	No order (0)	Order (1)
No order	41,933	**711**
Order	52	3,752

Evaluating Random Forest models: ROC curve and AUC





- The ROC curves are similar for both Random Forest models and also very similar to the logistic regression model
- The AUC score is slightly **higher for the tuned model** at 0.9926 versus 0.9919 for the default model
- **The logistic regression model has a higher AUC score than both Random Forest models at 0.9933**

Evaluating all models: key metrics

- The default Random Forest model resulted in the highest precision score which is what we are trying to optimise. However, it had a **lower accuracy and F1 score than both the logistic regression and tuned Random Forest models**

- The logistic regression model performed better than the tuned Random Forest model in terms of both the accuracy and F1 scores and **Pedal Power should therefore use the logistic regression model when predicting customer behaviour**

	Precision	Recall	Accuracy	F1 score
Calculation	TP / (TP + FP)	TP / (TP + FN)	(TP + TN) / (TP + FP + TN + FN)	2 * (Precision * Recall) / (Precision + Recall)
Logistic regression	**0.8451**	0.9853	0.984	0.9098
Random Forest (default)	**0.8470**	0.9735	0.9834	0.9058
Random Forest (tuned)	**0.8407**	0.9863	0.9836	0.9077

Conclusion

- A relatively small proportion (8%) of Pedal Power's website visitors placed an order in the last week suggesting this is a big ticket item which **requires some thought before purchase**
- A high number of customers have added a product to their wish list, suggesting these customers are interested in buying at some point. **Targeted advertising** may speed these purchases along
- May of the website visitors are returning customers, suggesting Pedal Power has done well to build **brand loyalty**
- Customers who viewed the checkout page had a **high propensity to buy**. This was backed up by correlation analysis which showed that viewing the checkout page was **positively correlated** with orders placed
- In order to predict propensity to buy going forward, Pedal Power are advised to **use the logistic regression** model – this produced a higher **AUC score, accuracy score and F1 score** than both the default and tuned Random Forest models
- The analysis could be improved by further tuning the Random Forest model, e.g. change hyper-parameters one at a time rather than all in one go or by training a different model or by trying **Grid Search Cross Validation** which would automatically choose the best hyper-parameters although this is **computationally expensive**

Thank you

Contact details:

Amy Birdee

amybirdee@gmail.com