Churn modelling

Alessandra Stagliano’

The problem

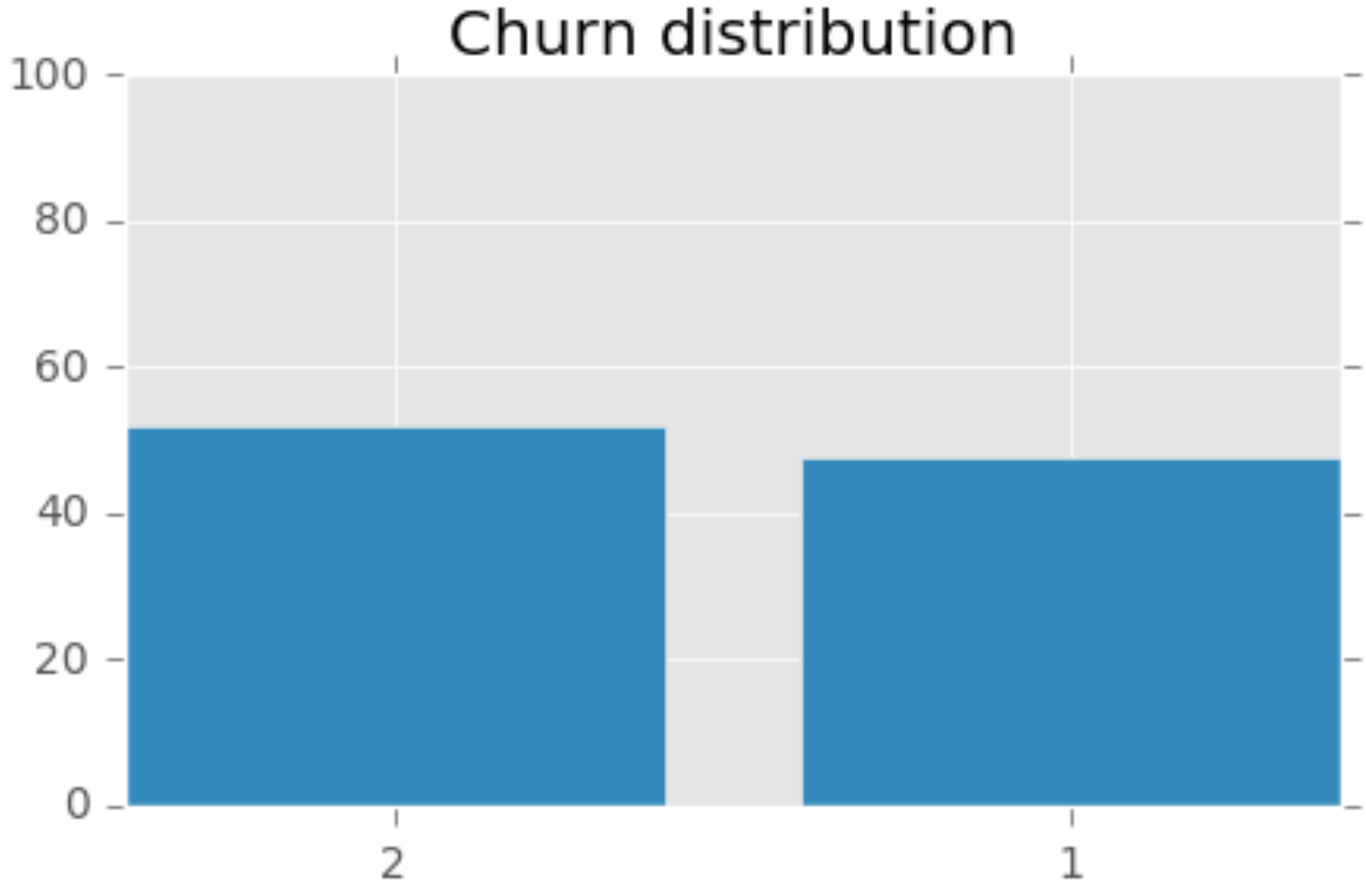
- Customer churn impedes business growth
- Being able to predict whether a customer is going to leave the service before they actually do is extremely valuable
- This information can be used to target potential churners, improving customer retention

The data

- For this short project, I was provided with data coming from different sources:
  - Customer data, including churn info, gender, country, account creation, year of birth, premiership
  - Receipts data, including quantity, date of purchase, price
  - Returns data, including reason for returning and action (reject, refund…)
  - Sessions data, including how clients interact with the page and the items

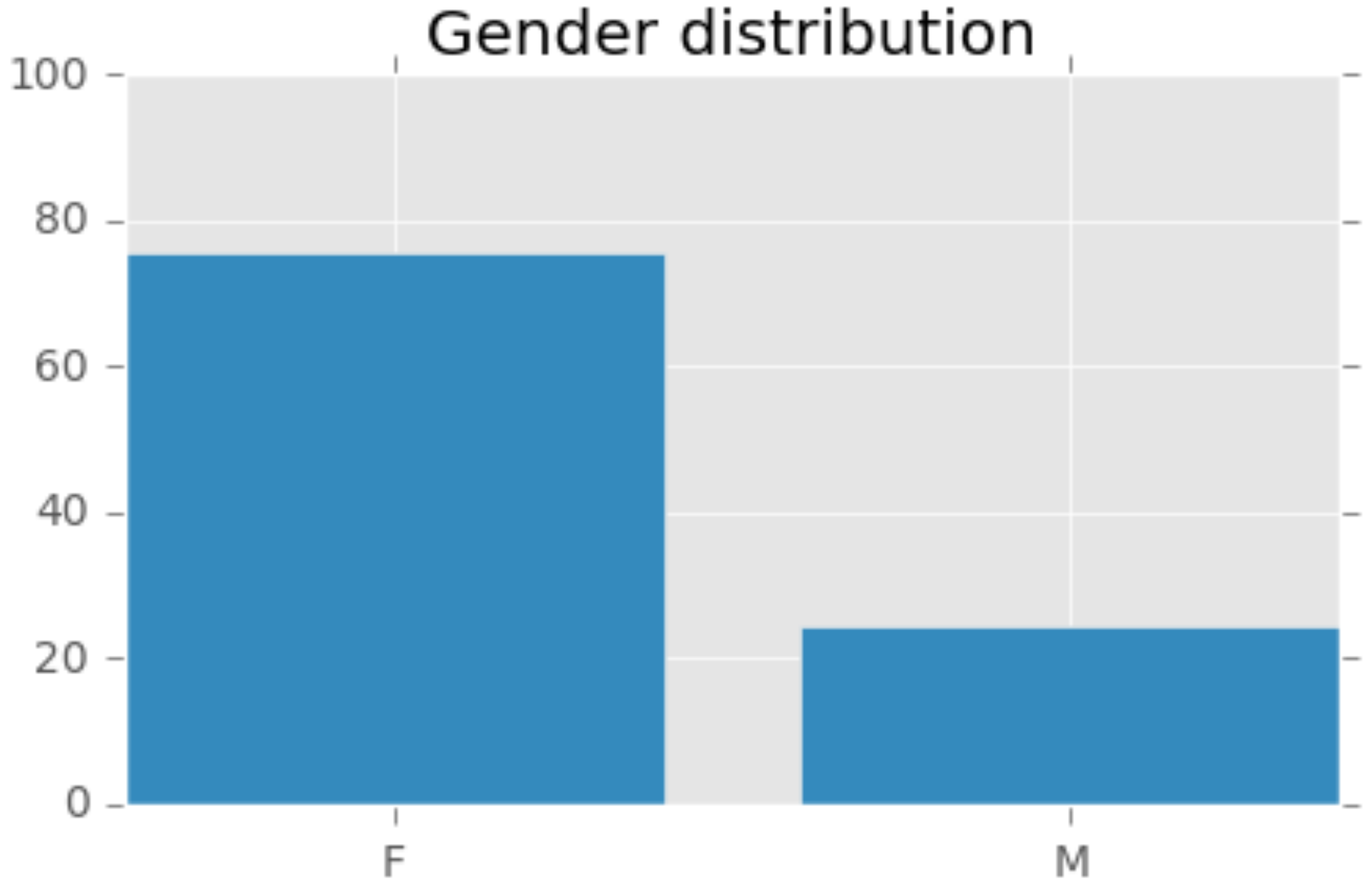
Exploration

- Is the dataset balanced?



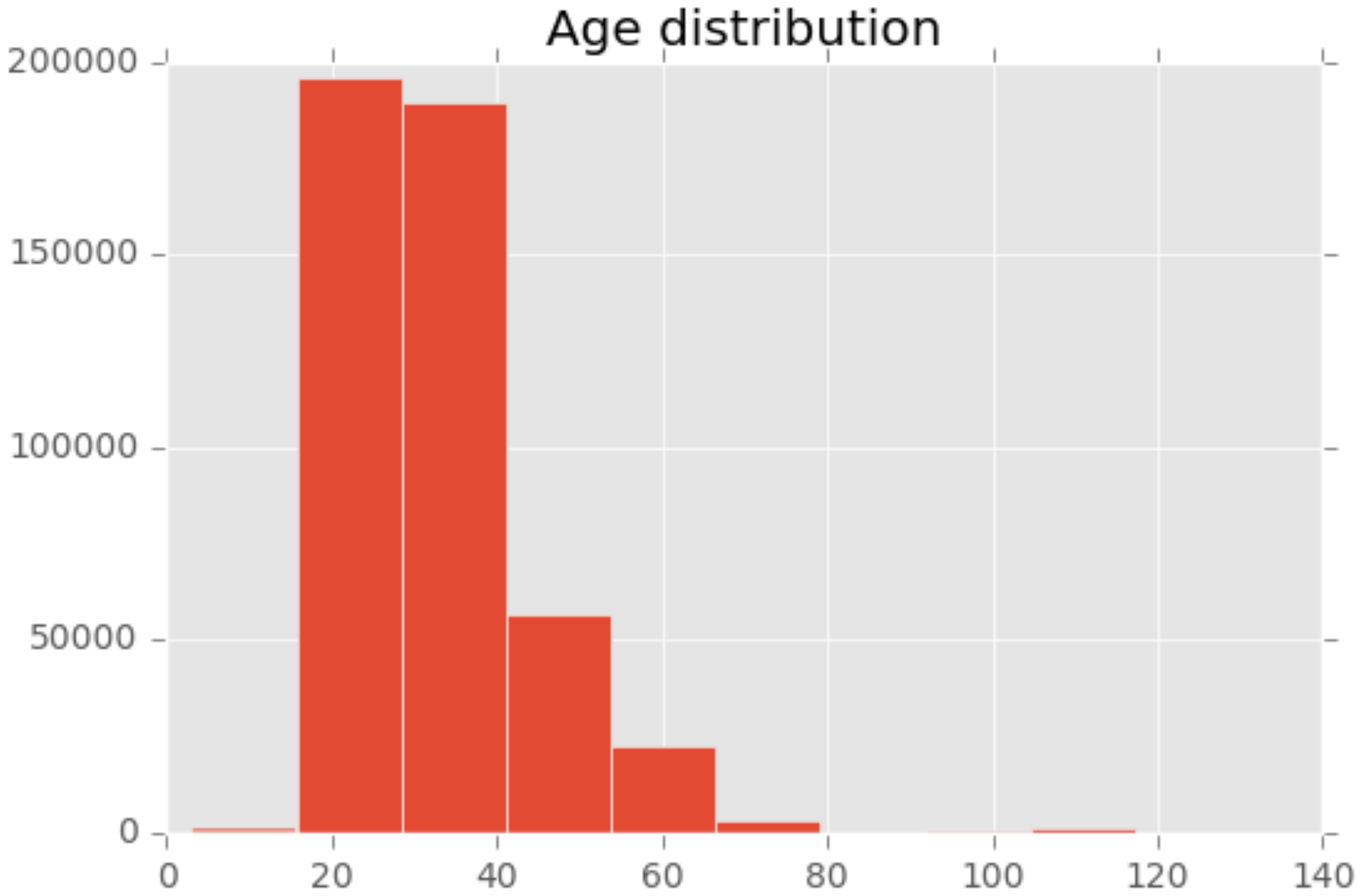
Exploration

- What is gender distribution?

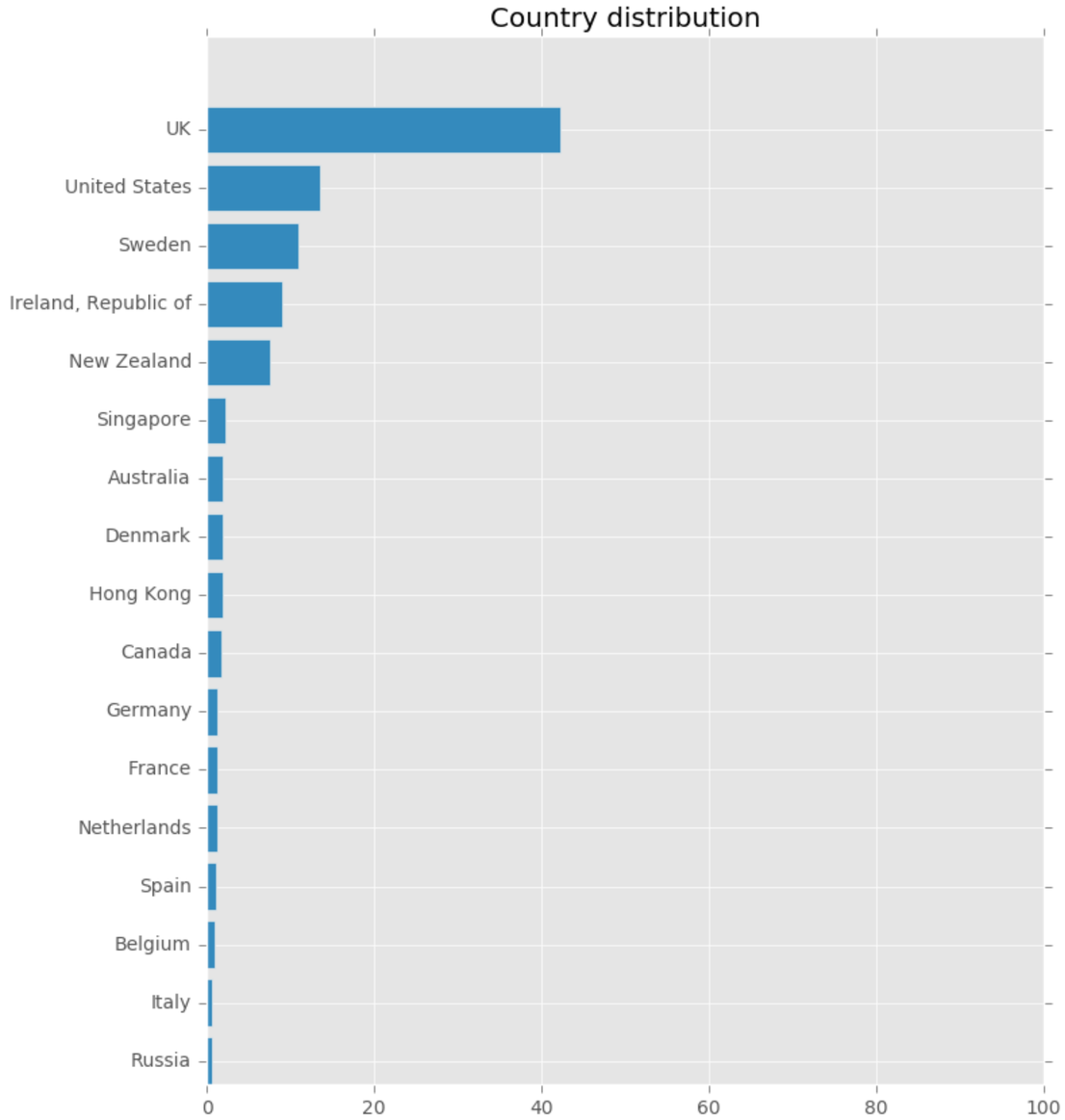


Exploration

- What is age distribution?



Exploration

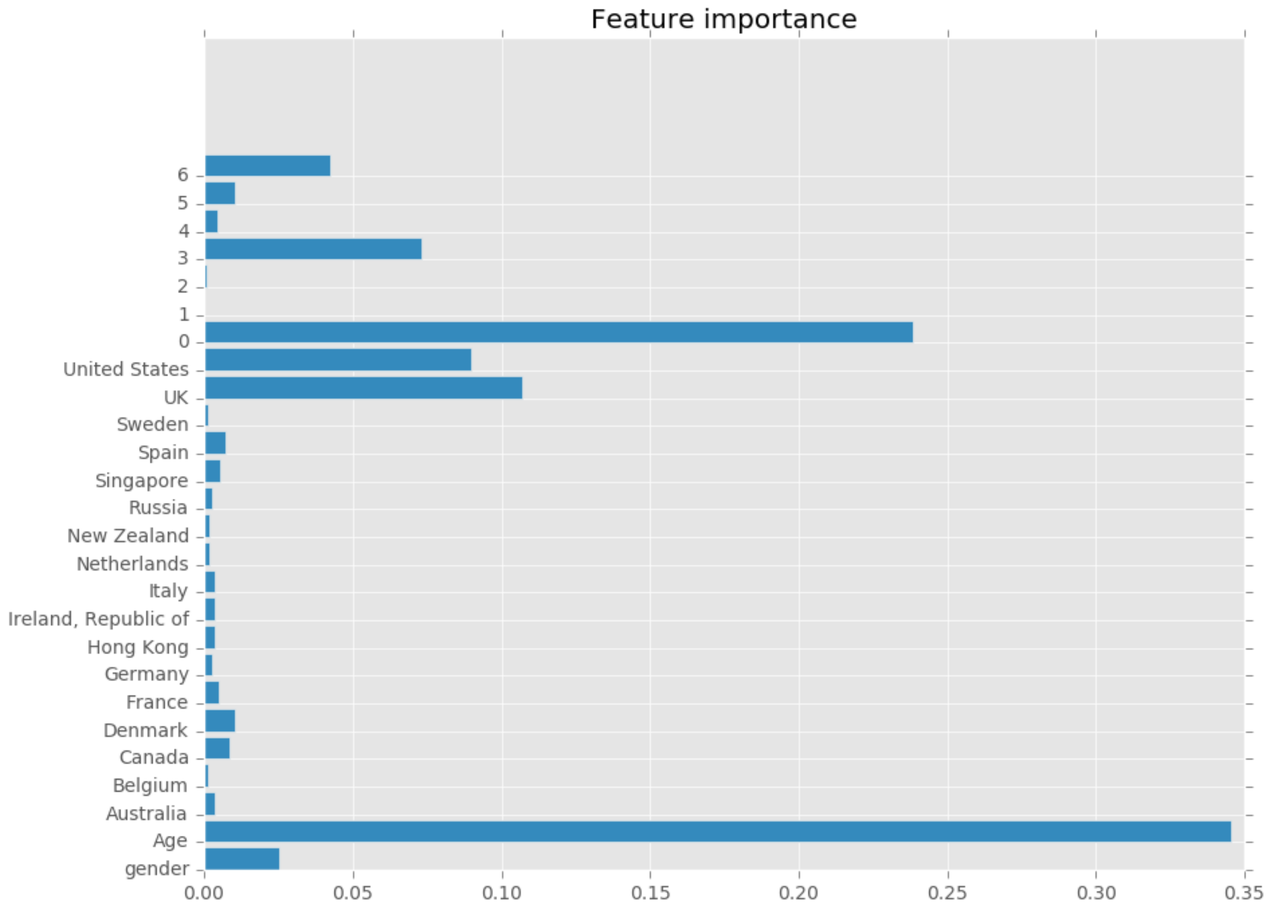


Benchmarking

- What can we say with really basic info about customers?
  - Gender
  - Age
  - Country of shipping
  - Premier

Benchmarking

- Model used: Random Forest
- Accuracy: 0.59
- Precision: 0.60
- Recall: 0.67

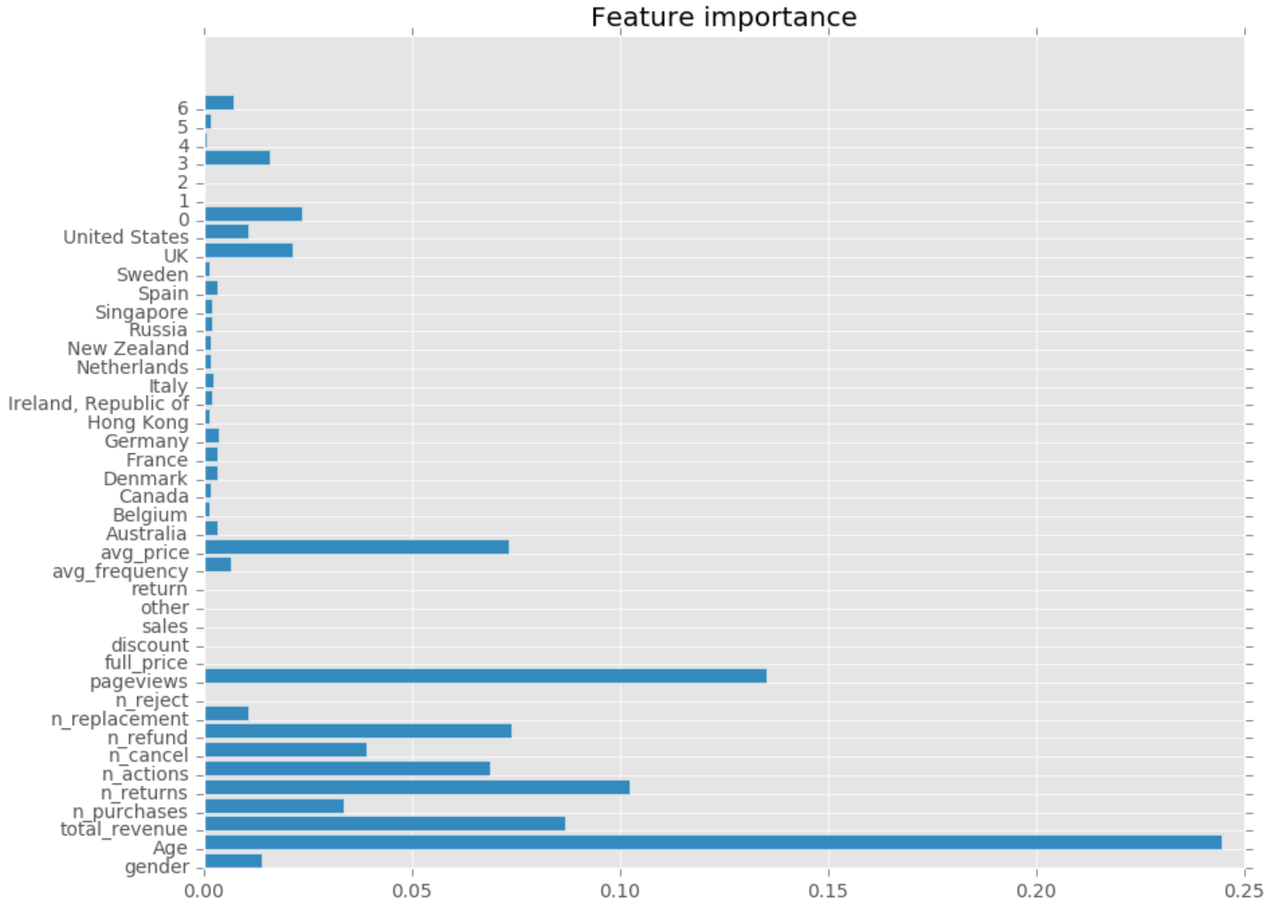


Building meaningful features

- Basic feature set plus:
  - Total revenue, # purchases, # returns, # of actions on website, # of cancelled orders, # refunds, # of replacements, # rejects, # page views, # items full price, # items discounted, # items on sale, # other purchases, average frequency of action
- All these features help in giving an idea of how customers behave

Model #1: Random Forests

- Accuracy: 0.60
- Precision: 0.63
- Recall: 0.61

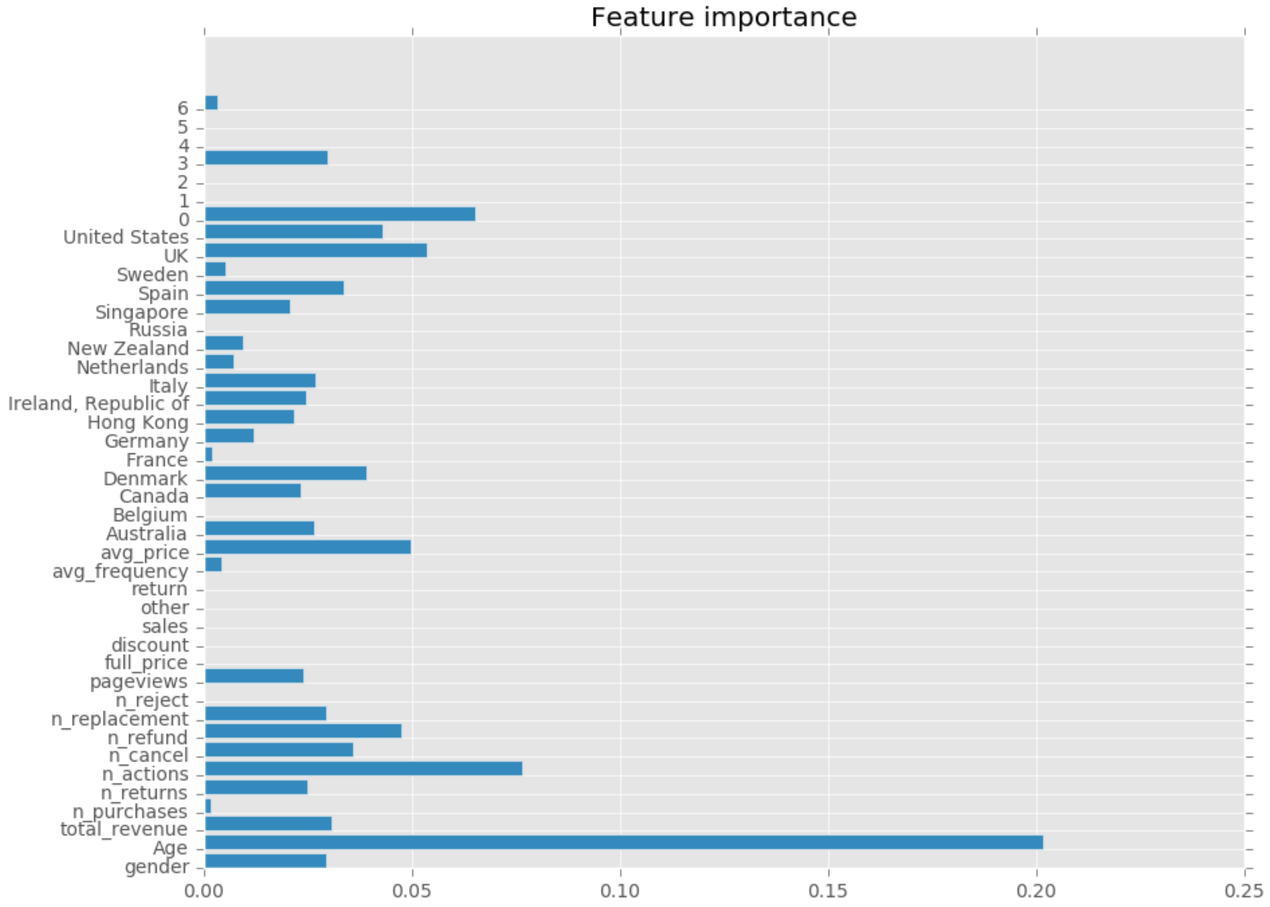


Model #2: Logistic Regression

- Accuracy: 0.61
- Precision: 0.63
- Recall: 0.63

Model #3: Gradient Boosting

- Accuracy: 0.63
- Precision: 0.65
- Recall: 0.65



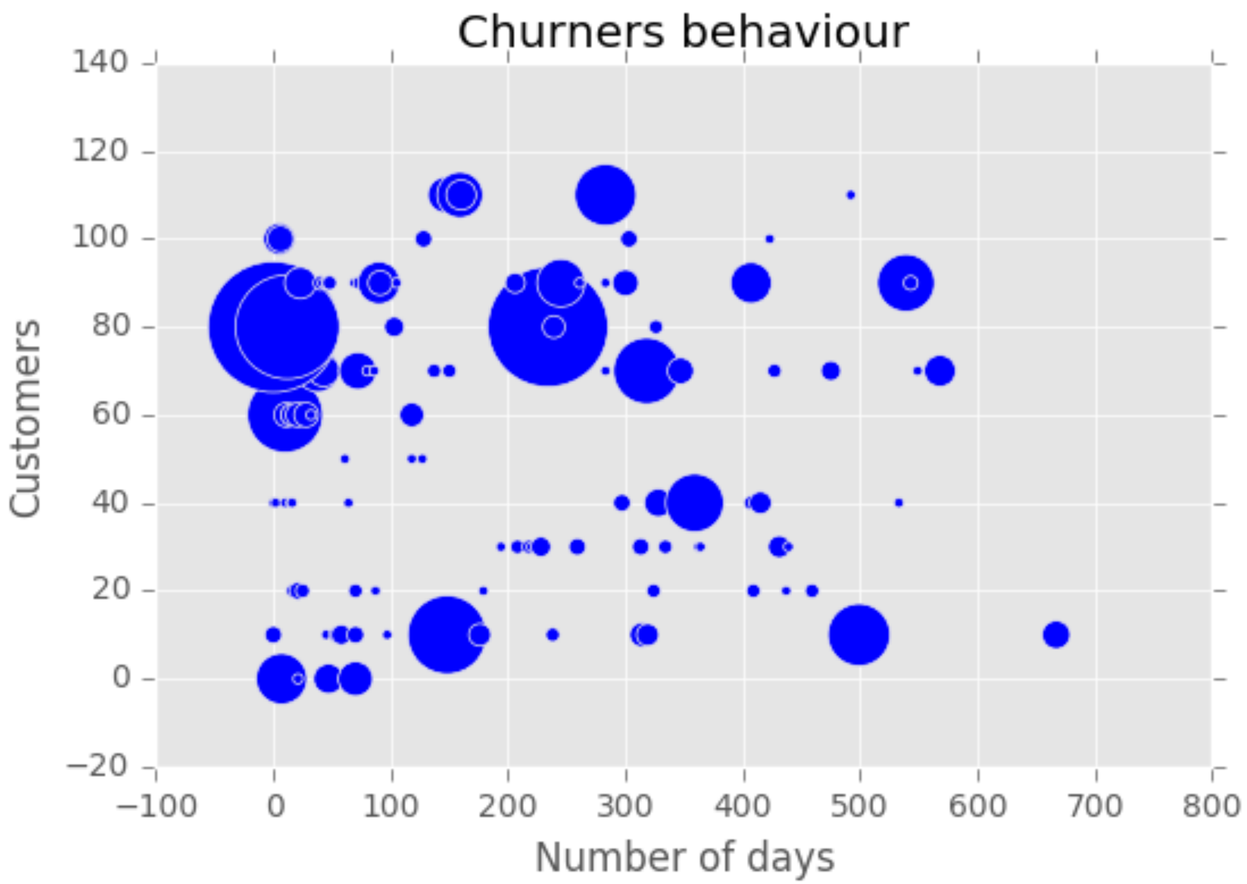
Considerations

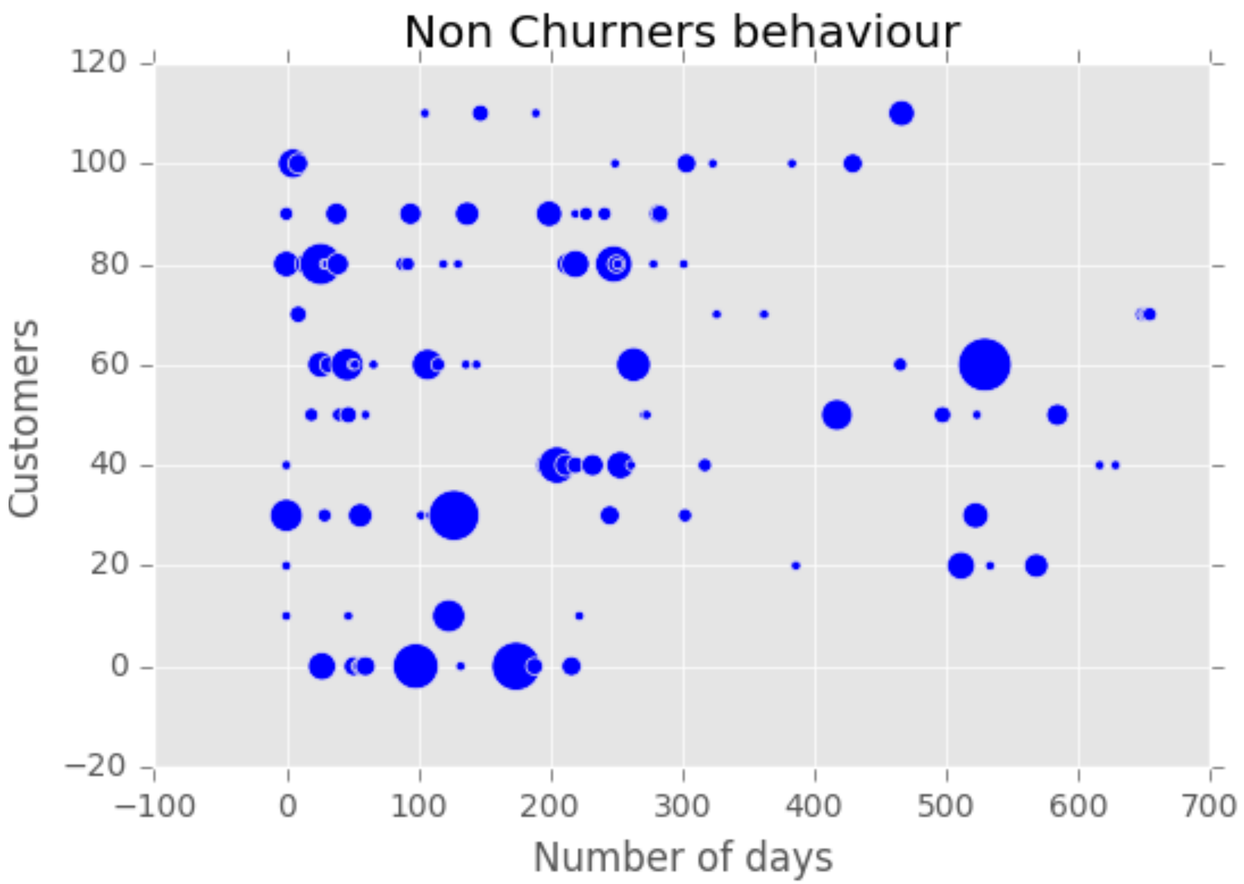
- An accuracy of +13% with respect to random choice means it could be possible to save ~300K pounds over ~500K customers
- As far as I know, there are about 15M active users using ASOS website (30 times the customer I have seen)
- If the proportion is the same (about 50% churners/non-churners) these 300K pounds could become ~9M pounds (without considering the fact that having more training data could lead to better predictive power, without adding anything)

If I had more time

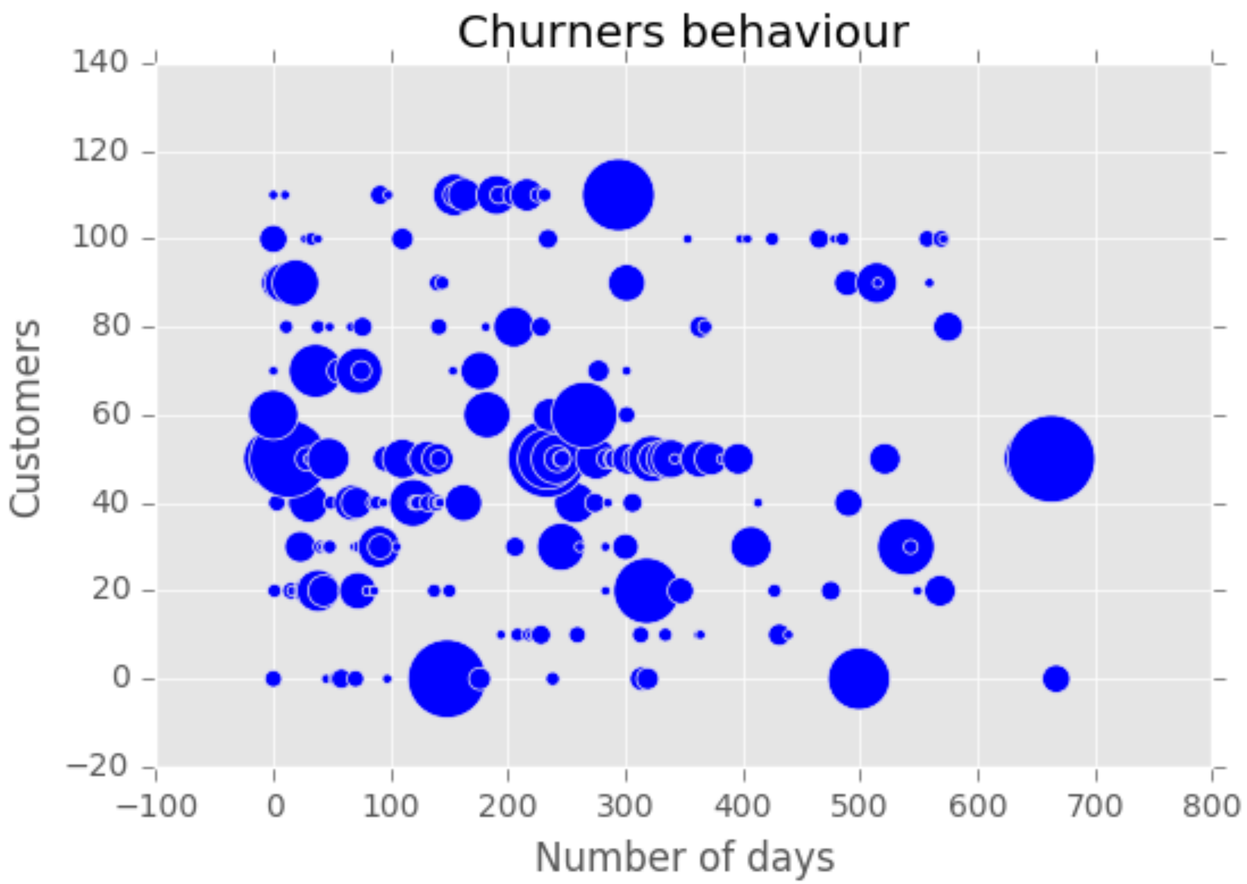
- Do a more rigorous cross validation / parameter tuning
- Spend more time selecting features
- Try other models

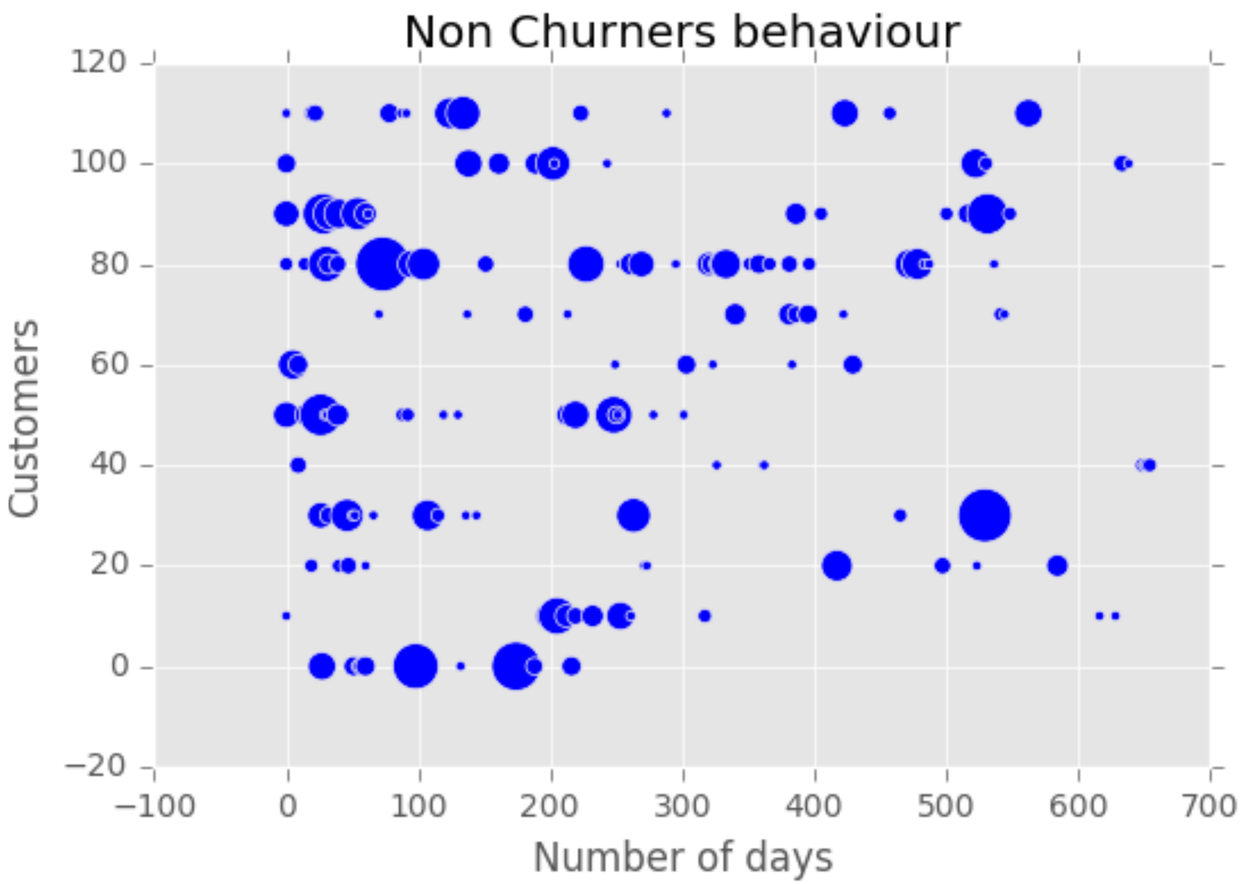
Quantitative insight - 10





Quantitative insight - 20





If I had more time

- Try to segment customers to see if there are groups with similar behaviour patterns
- Model clients interaction as a time series
- Change data representation: possibly apply similar embeddings to the ones used for the Lifetime Value work
    - this would allow the inclusion of additional information on the products viewed/purchased/returned

How can I help?

- Customer care data
  - if the text of email/messages is accessible: NLP to redirect messages to relevant people, understanding why people are churning, understanding what solutions worked and what not in case of complains
  - Would need tagged messages to know to which class they belong

How can I help?

- Stock optimisation:
  - Having in stock the minimum number of items would save a lot of money to the business. What is going to sell out? What should not be ordered, since it is unlikely it is going to sell?
  - Predictive model of sales, based on item features (from colour to price), seasonality, and possibly fashion expert knowledge

How can I help?

- Effective recommendations:
  - collaborative filtering can help in showing items to customers in the best way possible. Leveraging the “favourites” and purchase data it is possible to identify similar users and optimise the way images are shown to customers, leading to better conversion rates.

Appendix

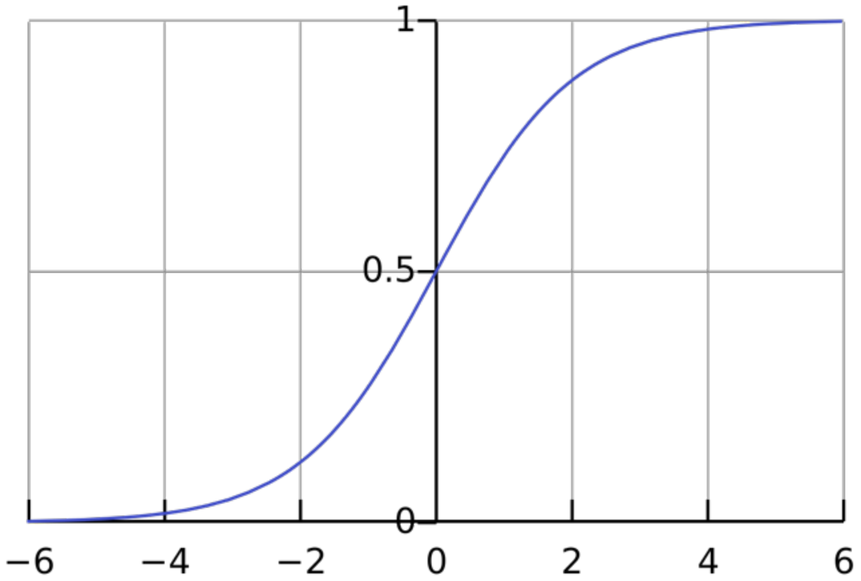
- Random Forest
- Logistic Regression
- Gradient Boosting Classifier

Random Forest

- RF are ensembles of decision trees
- Each tree of the ensemble is trained on a subset of the data and a subset of the features
- The label is assigned with a majority vote

Logistic Regression

- Regression model where the target variable is categorical
- The binary logistic model is used to estimate the probability of a binary response based on one or more features



$$F(x) = \frac{1}{1 + e^{-(\beta_0 + \beta_1 x)}}$$

Gradient Boosting Classifier

- Another ensemble model, based on trees
- Boosting is a technique that influences the way the data is subsampled
- While in RF each tree is independent of the others, in GB each tree is built on top of the one preceding it - the goal is to improve the performances over the misclassified data points