Postgraduate Diploma in Machine Learning & Artificial Intelligence

Capstone Project - Nick Heppermann

Background

Company: Nuveen

Invests in the growth of businesses, real estate, farmland, forests and infrastructure while building lifetime relationships with clients from all over the globe.

Context

- Nuveen **sells financial products** to its clients (**financial professionals and individual investors**)
- Nuveen's **marketing department needs intelligence** on its clients to take appropriate marketing actions
- **Provide** Nuveen **marketing intelligence** so Nuveen can **effectively deploy** their **marketing efforts**

Objectives - help Nuveen gather marketing intelligence

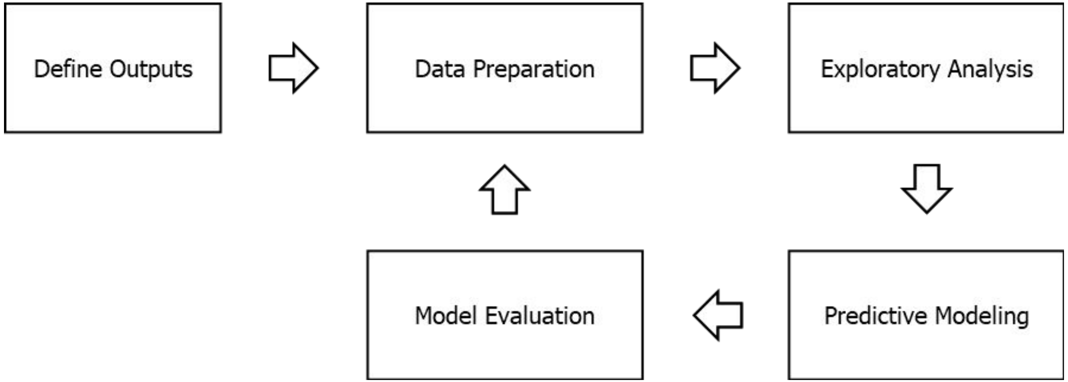
Predict client’s next year sales and whether client will add a fund so marketing team can:

- Assess client potential
- Prioritize high value clients
- Perform appropriate marketing tasks

Q: Why data science?

A: A strongly predictive data pipeline could serve as a tool to help the marketing department identify high value clients

Approach - overview of data analysis effort



- Outputs
  - Sales prediction
  - New fund added prediction
- Data preparation
- Exploratory analysis

- Predictive modeling
  - Sales → regression model
  - New fund added → classification model
- Model evaluation
  - R^2 score and lift over average

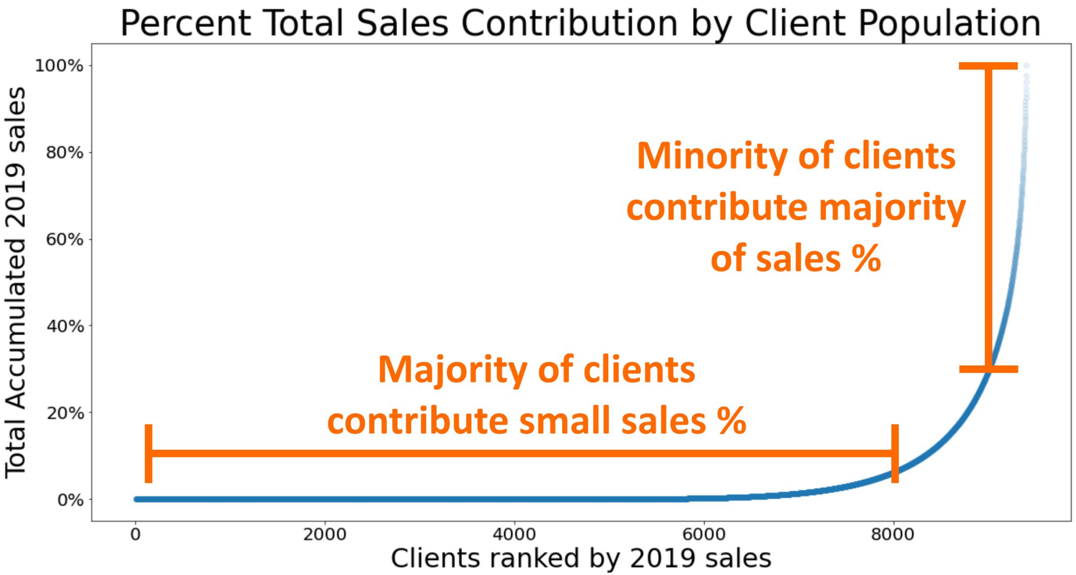
Approach - data preparation (both sales regression and classification)

- Models were trained from data in provided spreadsheets (no outside data was used)
- Small percentage of samples were removed due to lack of data or nonsensical data
  - Omitted 585 clients (~2% of samples) due to no numeric data
  - Omitted 17 clients (<<1% of samples) due to negative sales (2018 or 2019) or positive redemptions
- Missing values set to zero

See Appendix A for data preparation specifics

Predicting Sales

Approach - meet your data (sales regression)

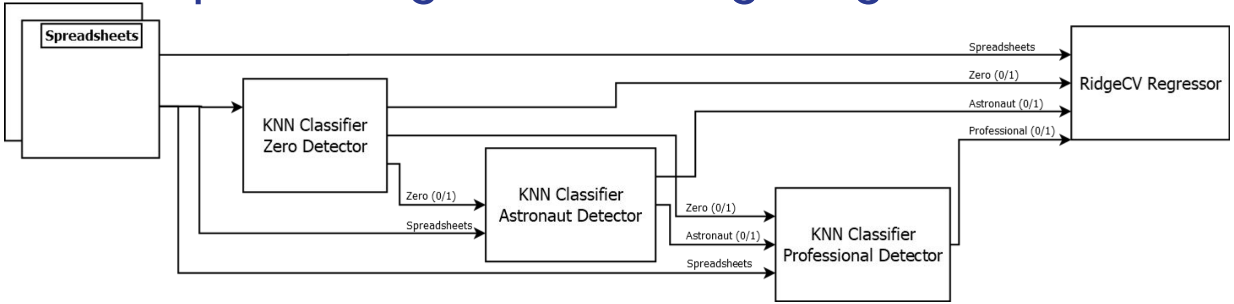


Minority of clients contribute majority of sales %

Majority of clients contribute small sales %

- 2019 client sales numbers are extremely top heavy
  - Majority of clients contribute minority of sales
- Goal: Identify the high performers, and weed out those that contribute nothing

Approach - predicting sales through regression



- Overall model design relied on two main philosophies
  - Separate clients into sales buckets (coarse prediction)
  - Take sales bucket predictions and spreadsheet data as inputs to continuous prediction model (fine grain prediction)
- Sales buckets used in coarse prediction

'Zeros' bucket (2019 sales < $1)

'Avg Joes/Janes' bucket ($1 <= 2019 sales < $368,000) not modeled

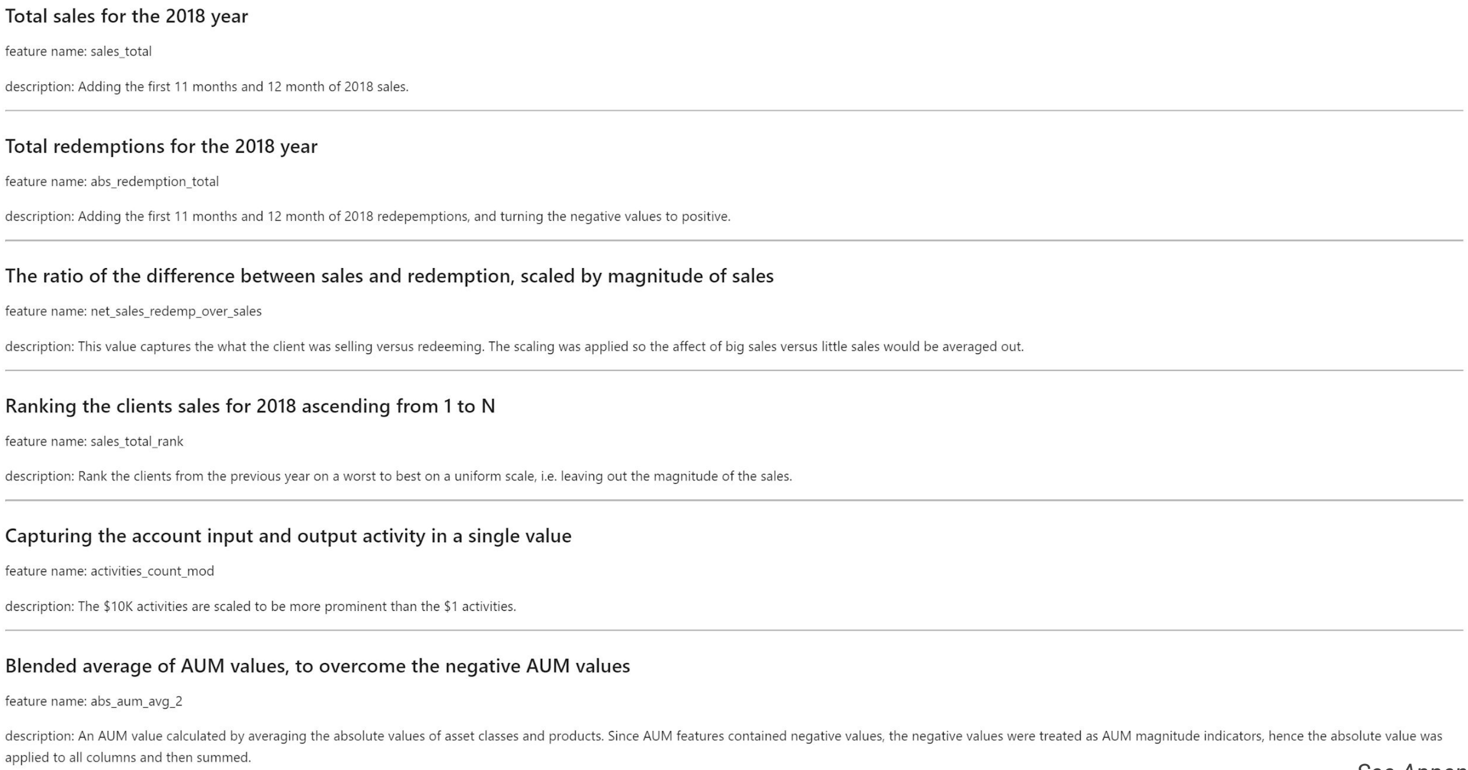
'Professionals' bucket ($368,000 <= 2019 sales < $3,700,000)

'Astronaut' bucket ($3,700,000 <= 2019 sales)

See Appendix B for model design specifics

Approach - creating custom data features

These custom data features prominently influenced the models



Total sales for the 2018 year

feature name: sales_total

description: Adding the first 11 months and 12 month of 2018 sales.

Total redemptions for the 2018 year

feature name: abs_redemption_total

description: Adding the first 11 months and 12 month of 2018 redepemptions, and turning the negative values to positive.

The ratio of the difference between sales and redemption, scaled by magnitude of sales

feature name: net_sales_redemp_over_sales

description: This value captures the what the client was selling versus redeeming. The scaling was applied so the affect of big sales versus little sales would be averaged out.

Ranking the clients sales for 2018 ascending from 1 to N

feature name: sales_total_rank

description: Rank the clients from the previous year on a worst to best on a uniform scale, i.e. leaving out the magnitude of the sales.

Capturing the account input and output activity in a single value

feature name: activities_count_mod

description: The $10K activities are scaled to be more prominent than the $1 activities.

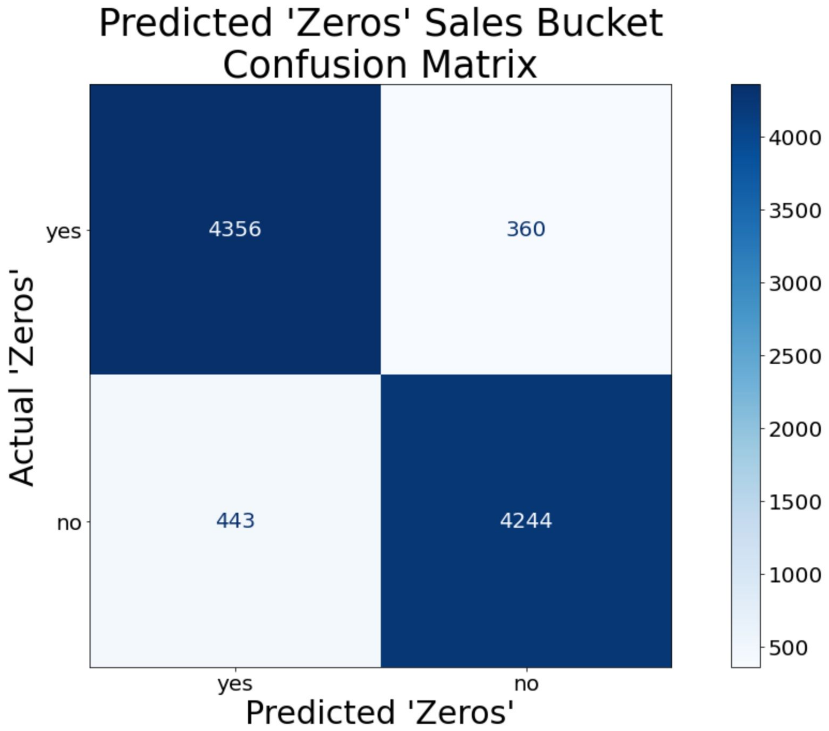
Blended average of AUM values, to overcome the negative AUM values

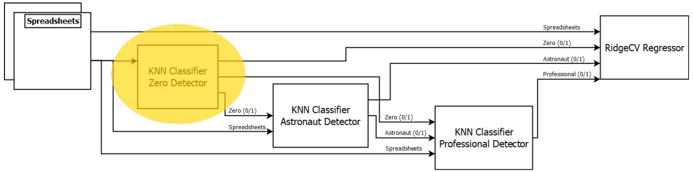
feature name: abs_aum_avg_2

description: An AUM value calculated by averaging the absolute values of asset classes and products. Since AUM features contained negative values, the negative values were treated as AUM magnitude indicators, hence the absolute value was applied to all columns and then summed.

See Appendix C for variable equations

Results - predicting sales through sales buckets - zeros





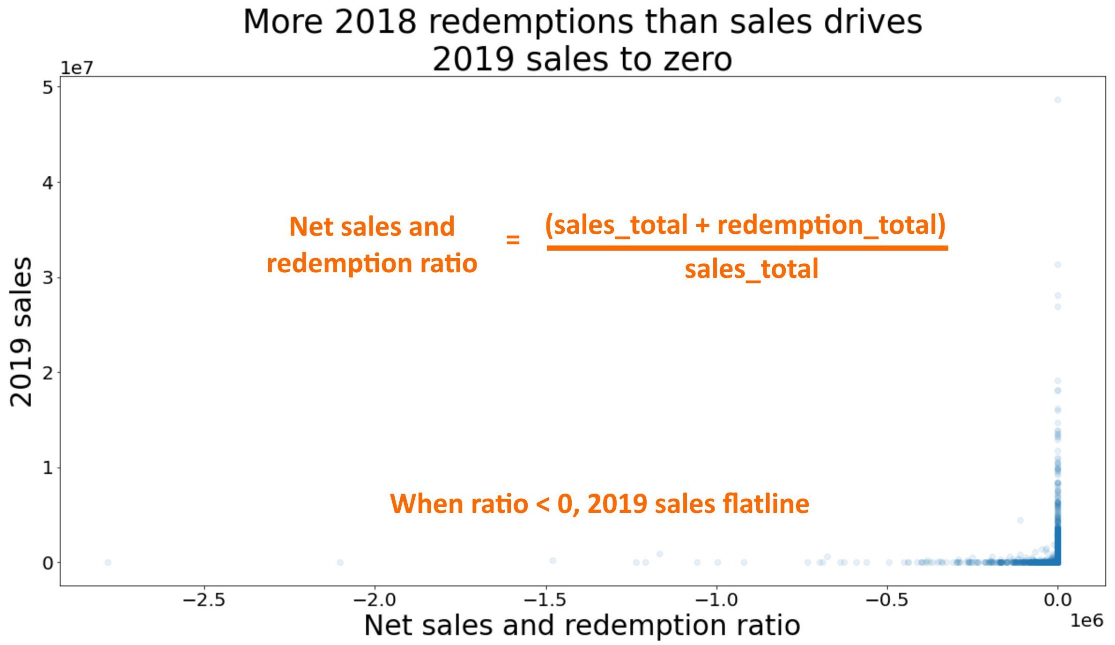
- Accurate classification of ‘zeros’, predicted clients that contribute little
  - All ‘zeros’ had no sales in final months, classifier filtered those clients out as a rule
- Predicted ‘zeros’ only accounted for 4.55% of total 2019 sales → not much loss due to misclassification

See Appendix F for classifier scores

Results - identify zeros by the numbers

- All ‘zeros’ had no sales in the final month, so if final month had sales → not a ‘zero’

- Large redemption totals more likely to client being a ‘zero’
  - Approach large redemptions with:
    - retention strategies
    - associate list if retiring

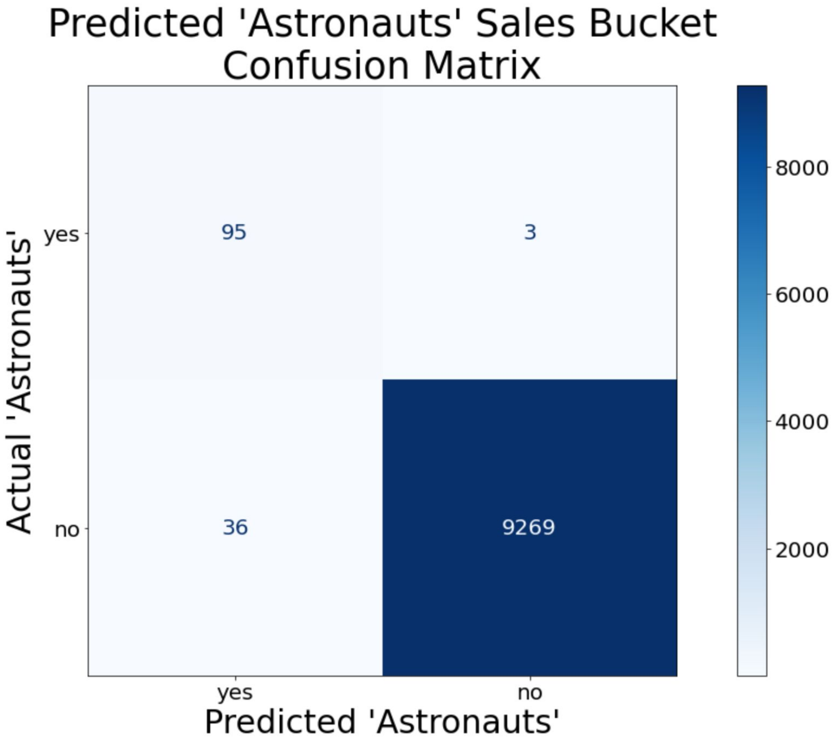


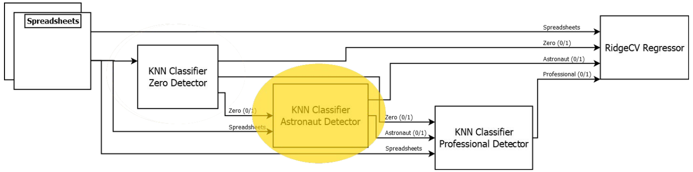
More 2018 redemptions than sales drives 2019 sales to zero

$$\text{Net sales and redemption ratio} = \frac{(\text{sales\_total} + \text{redemption\_total})}{\text{sales\_total}}$$

When ratio < 0, 2019 sales flatline

Results - predicting sales through sales buckets - astronauts



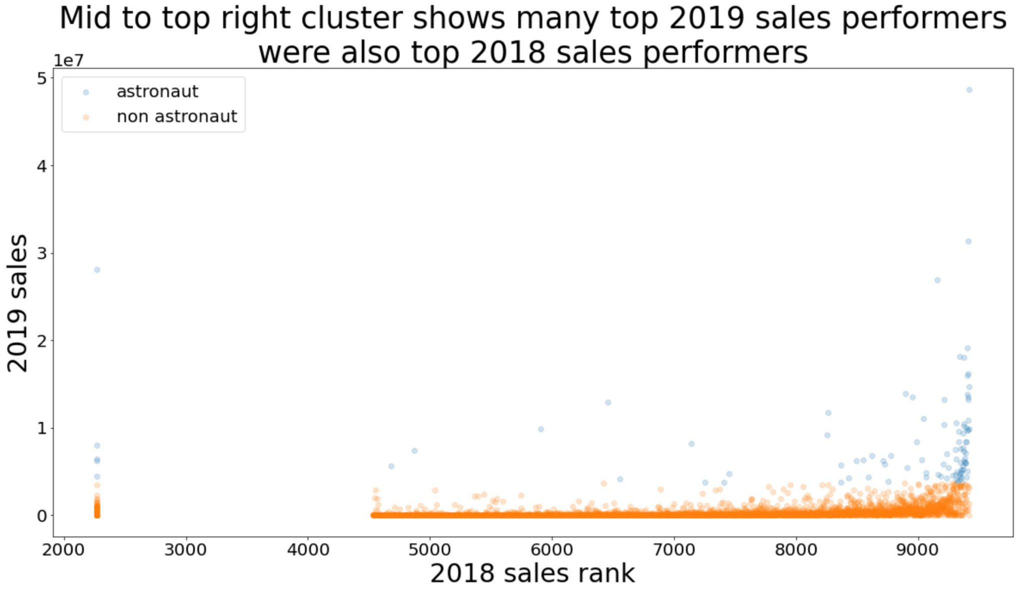


- Accurate classification of ‘astronauts’, predicted clients that contributed a lot
- 95 of 98 ‘astronauts’ identified ensures bulk of sales will be retained
- Classifier was greedy, classified 36 false positives
  - False positives still averaged 2019 sales of ~$1M → still high value targets

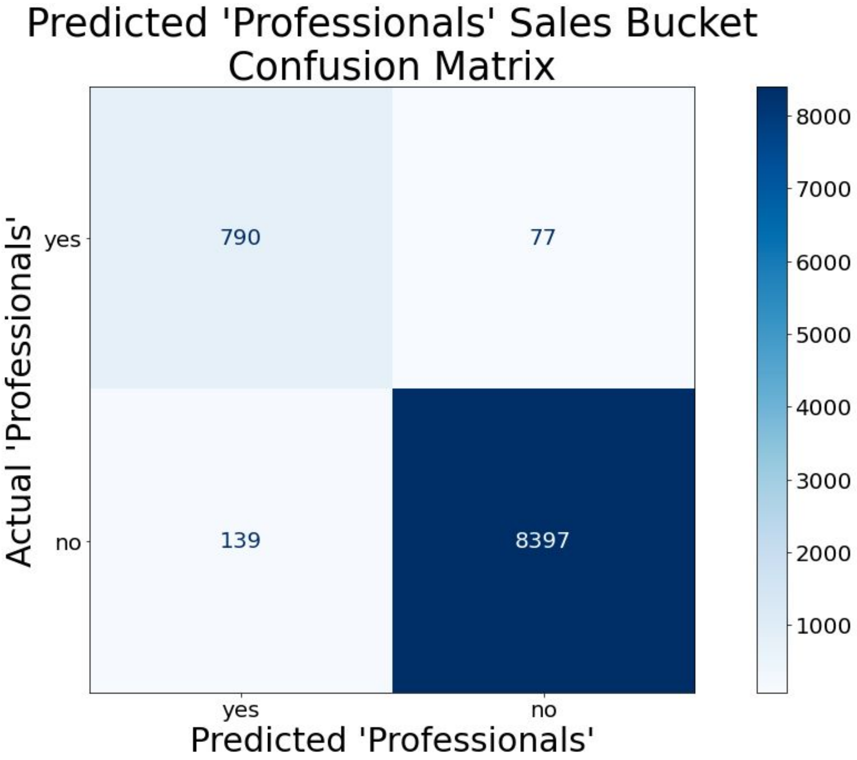
See Appendix G for classifier scores

Results - identify astronauts by the numbers

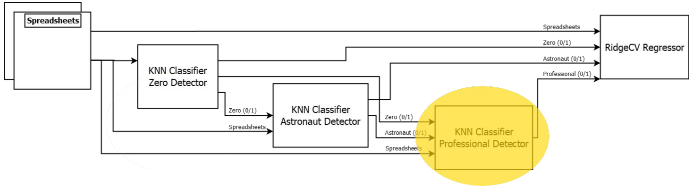
- Had lots of account activity,
  so high counts in sales & redemptions
- Ranked high in 2018 sales
- Had high AUM
- Influencing channels:
  - Less likely to be Independent Dealer
    - 48% astros, 69% overall
  - More likely to be National Broker-Dealer
    - 45% astros, 21% overall
- Influencing firms:
  - More likely Merrill Lynch
    - 20% astros, 8% overall
  - More likely Morgan Stanley Wealth Mgmt
    - 17% astros, 8% overall



Results - predicting sales through sales buckets - professionals



Predicted 'Professionals' Sales Bucket Confusion Matrix

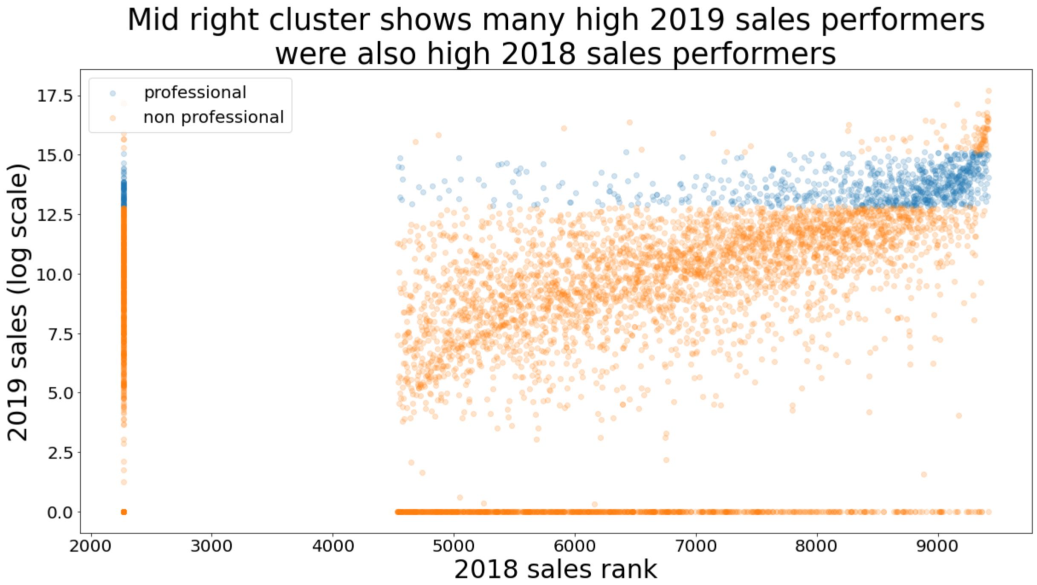


- Somewhat accurate classification of ‘heros’
- Classifier missed 77 (9%) ‘professionals’
  - True negatives averaged 2019 sales of ~$860k → ouch
- Classifier was greedy, classified 139 false positives
  - False positives still averaged 2019 sales of ~$200k → still high value targets

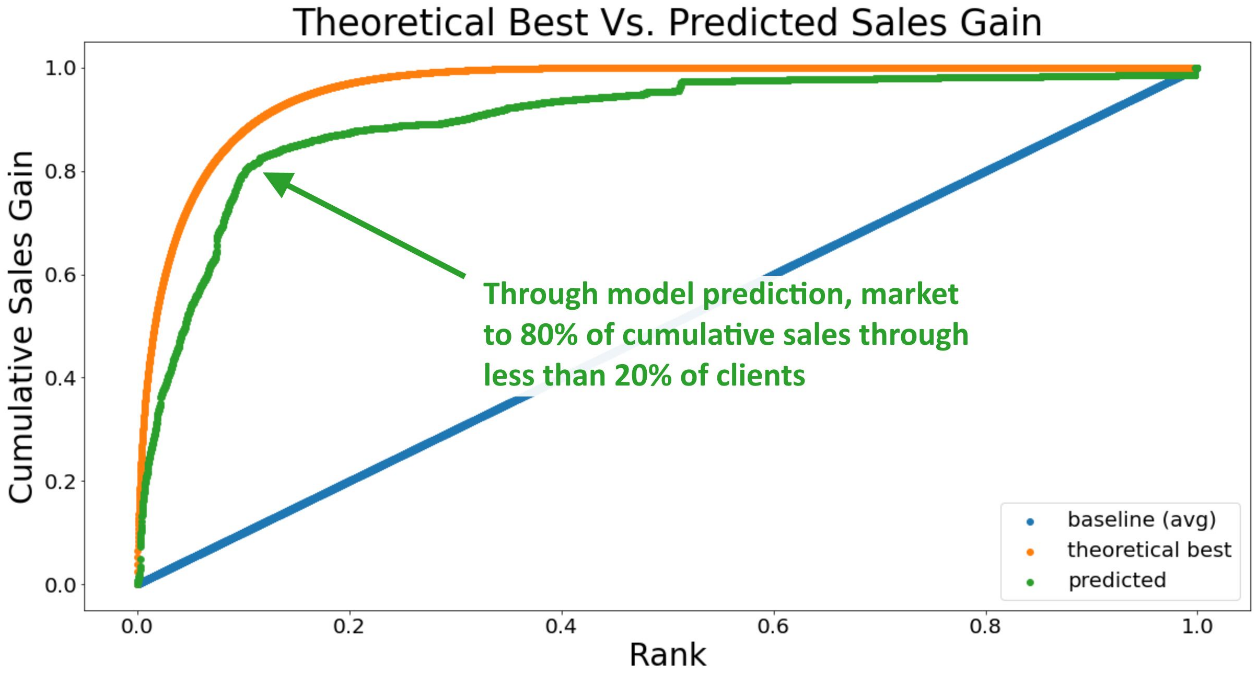
See Appendix H for classifier scores

Results - identify professionals by the numbers

- Had lots of account activity, so high counts in sales & redemptions
- Ranked high in 2018 sales
- Had high AUM
- Influencing channels:
  - Less likely Independent Dealer
    - 52% astros, 69% overall
  - More likely National Broker-Dealer
    - 40% astros, 21% overall
- Influencing firms:
  - More likely Merrill Lynch
    - 17% astros, 8% overall
  - More likely Morgan Stanley Wealth Mgmt
    - 16% astros, 8% overall



Results - regression - reviewing overall performance



Theoretical Best Vs. Predicted Sales Gain

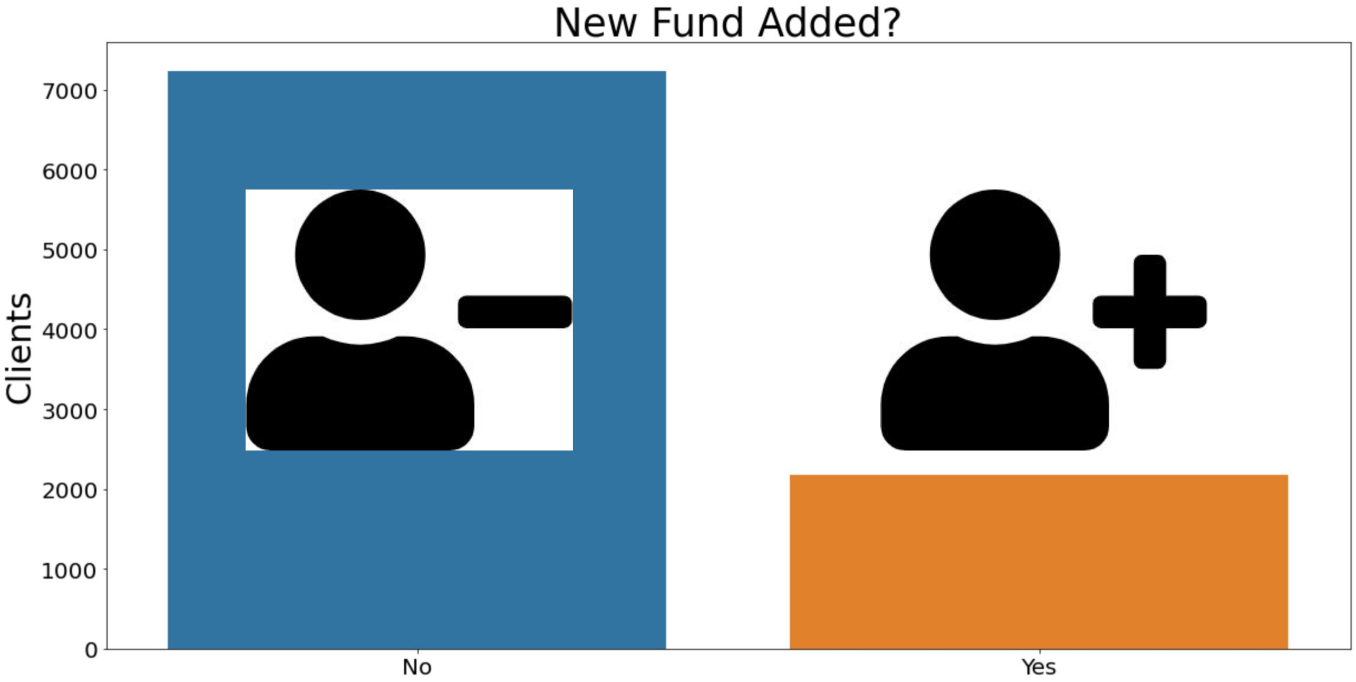
Through model prediction, market to 80% of cumulative sales through less than 20% of clients

- Predicted cumulative gains mostly tracks theoretical best curve
    - Top sales clients could be targeted with reduced effort/expense
- Target top 20% of clients to market to 80% of predicted sales

See Appendix I for regressor scores

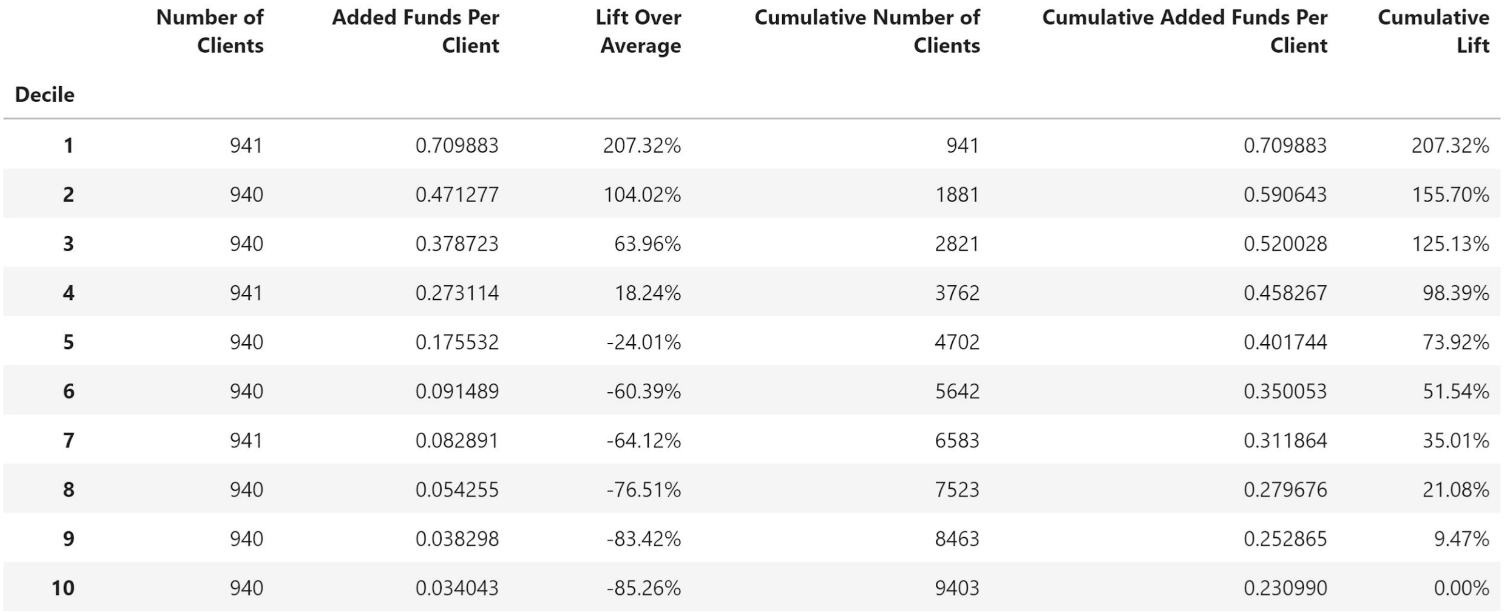
Predicting Added Funds

Approach - meet your data (classification)



- Only 23% of clients will add a fund

Results - targeting the most likely (classification)



	Number of Clients	Added Funds Per Client	Lift Over Average	Cumulative Number of Clients	Cumulative Added Funds Per Client	Cumulative Lift
Decile						
1	941	0.709883	207.32%	941	0.709883	207.32%
2	940	0.471277	104.02%	1881	0.590643	155.70%
3	940	0.378723	63.96%	2821	0.520028	125.13%
4	941	0.273114	18.24%	3762	0.458267	98.39%
5	940	0.175532	-24.01%	4702	0.401744	73.92%
6	940	0.091489	-60.39%	5642	0.350053	51.54%
7	941	0.082891	-64.12%	6583	0.311864	35.01%
8	940	0.054255	-76.51%	7523	0.279676	21.08%
9	940	0.038298	-83.42%	8463	0.252865	9.47%
10	940	0.034043	-85.26%	9403	0.230990	0.00%

Target clients in deciles 1 through 3 to achieve optimal marketing efficiency

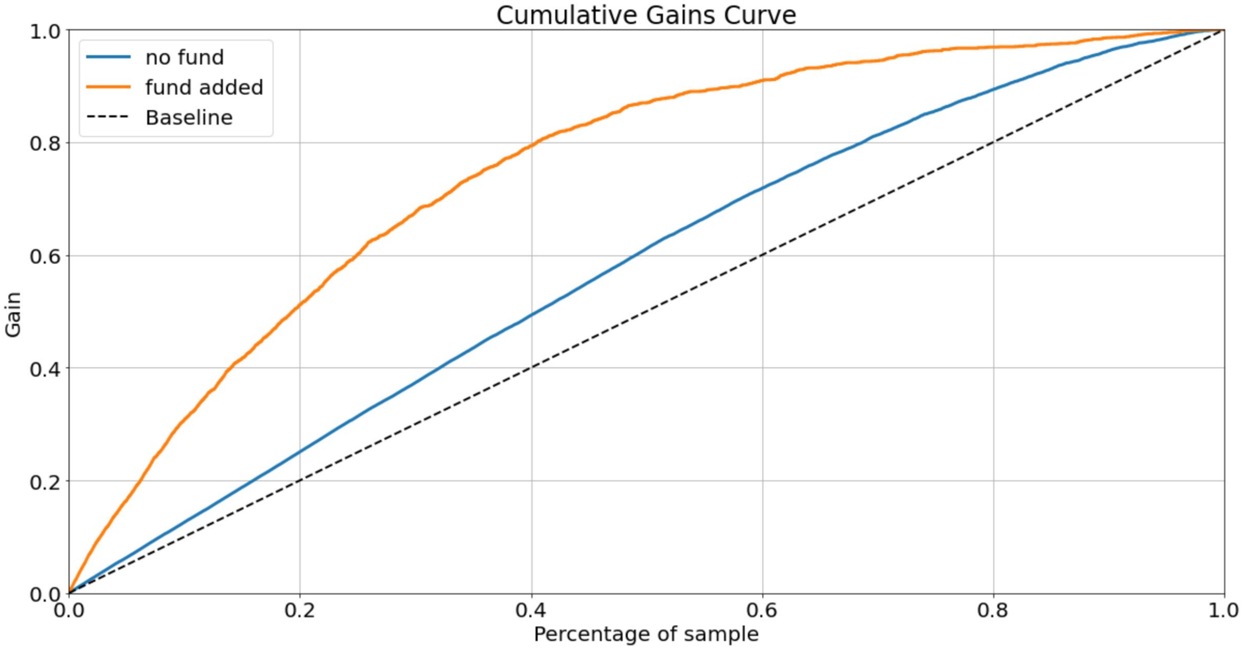
See Appendix E for classifier scores

Results - identifying clients that add funds through numbers

- More active in sales and redemptions
- More likely to not be in the ‘zeros’ sales bucket
  - 87% of clients that added funds were predicted to not be ‘zeros’
  - Clients with zero sales did not add funds
- Influencing channel - more likely to be National Broker-Dealer
  - 29% of the added fund group, vs 18% of the not added fund group
- Influencing sub-channels
  - Less likely to be IBD
    - 58% of the added fund group, vs 67% of the not added fund group
  - More likely to be NACS
    - 40% of the added fund group, vs 29% of the not added fund group

- Influencing firms - more likely to be in a large firm
  - Edward Jones
    - 12% of the added fund group, vs 8% of the not added fund group
  - Merrill Lynch
    - 12% of the added fund group, vs 7% of the not added fund group
  - Morgan Stanley Wealth Management
    - 12% of the added fund group, vs 7% of the not added fund group
- Influencing firms - less likely to be in a small firm
    - 20% of the added fund group, vs 38% of the not added fund group

Results - targeting the most likely (classification)



Deciles 1 through 4 have higher than average clients with funds added

The End

I hope you enjoyed the presentation.

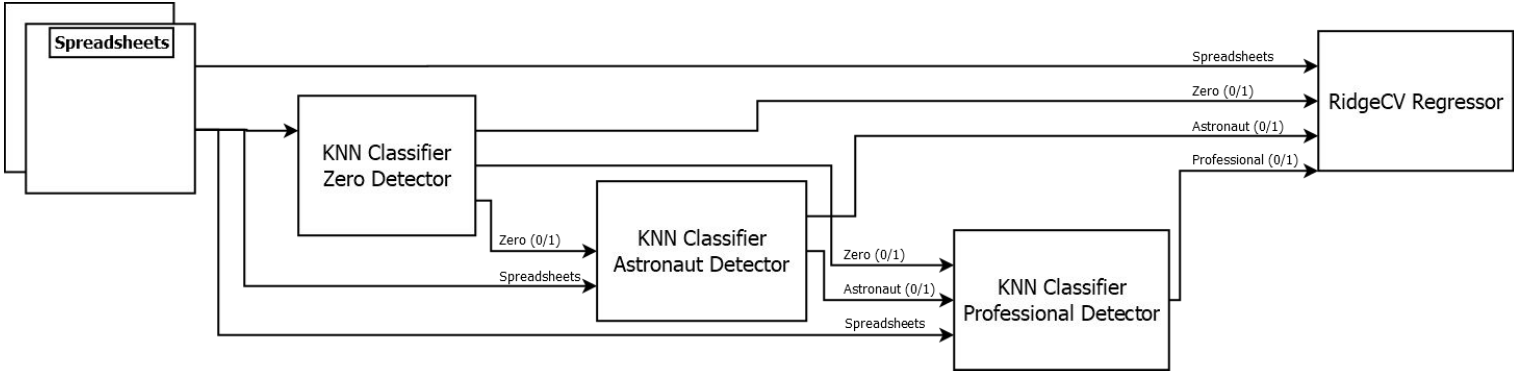
I enjoyed the class -
thank you for providing it.

Appendix follows →

Appendix A - data preparation

- Transaction Data.xlsx ('Transactions18' sheet)
  - Contains numeric cient data
  - Blank values filled with 0
  - Omitted 585 clients (~2% of samples) due to no numeric data
  - Omitted 17 clients (<<1% of samples) due to negative sales (2018 or 2019) or positive redemptions
  - Missing values set to zero
- Transaction Data.xlsx ('Transactions19' sheet)
  - Contains output values for sales regression and new fund added
  - Missing values set to zero
  - New fund added converted to mapped to 0 or 1
- Firm Information.xlsx ('Rep summary' sheet)
  - Contains client firm information
- Firm Information.xlsx ('Asset fund summary' sheet) - not used

Appendix B - predicting sales through regression

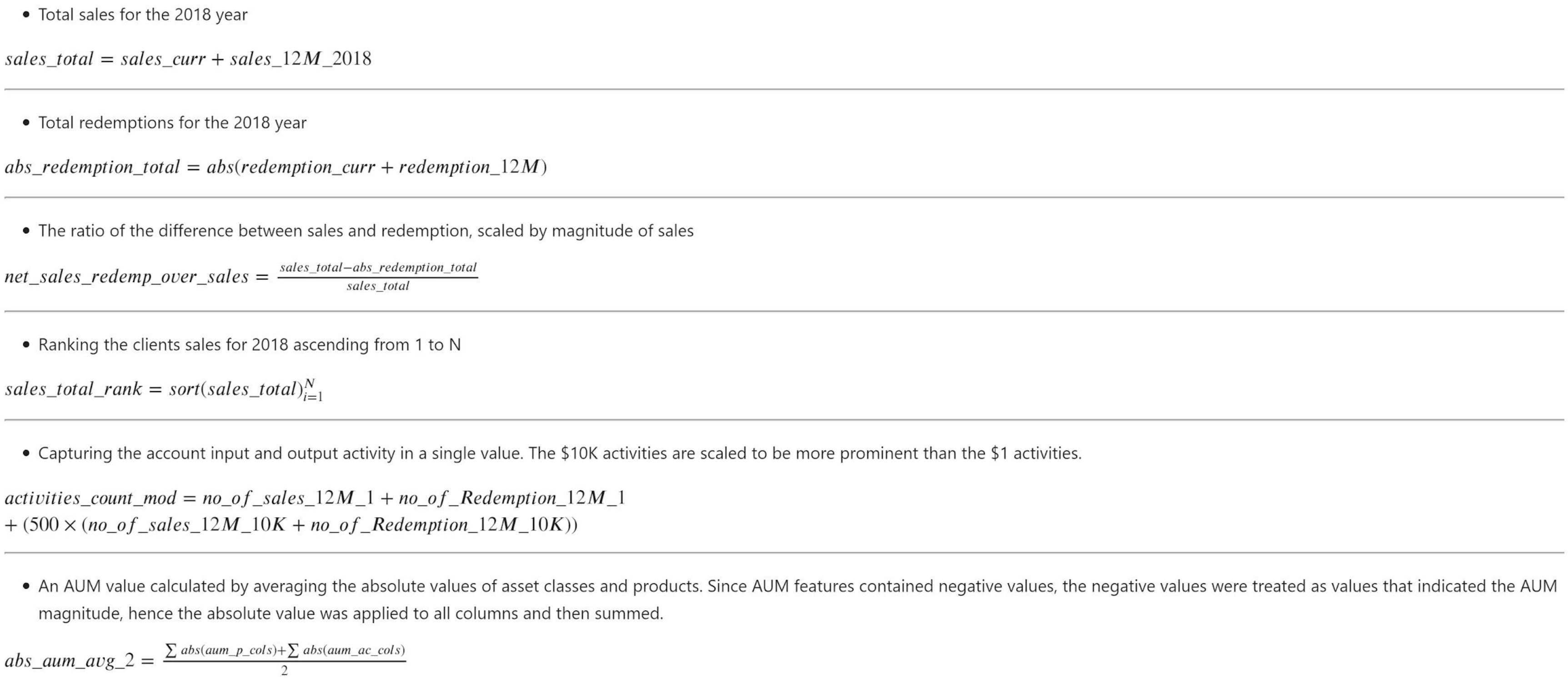


- 4 models in total used to predict sales
- 3 KNN classifiers - each classifying a different sales buckets,

  predicted values serve as inputs to final model along with spreadsheet data
- RidgeCV regressor - final model
  - Regression model chosen to predict continuous output value

Appendix C - creating custom data features

These are the equations for the custom data features



- Total sales for the 2018 year

$$sales\_total = sales\_curr + sales\_12M\_2018$$

- Total redemptions for the 2018 year

$$abs\_redemption\_total = abs(redemption\_curr + redemption\_12M)$$

- The ratio of the difference between sales and redemption, scaled by magnitude of sales

$$net\_sales\_redemp\_over\_sales = \frac{sales\_total - abs\_redemption\_total}{sales\_total}$$

- Ranking the clients sales for 2018 ascending from 1 to N

$$sales\_total\_rank = sort(sales\_total)_{i=1}^{N}$$

- Capturing the account input and output activity in a single value. The $10K activities are scaled to be more prominent than the $1 activities.

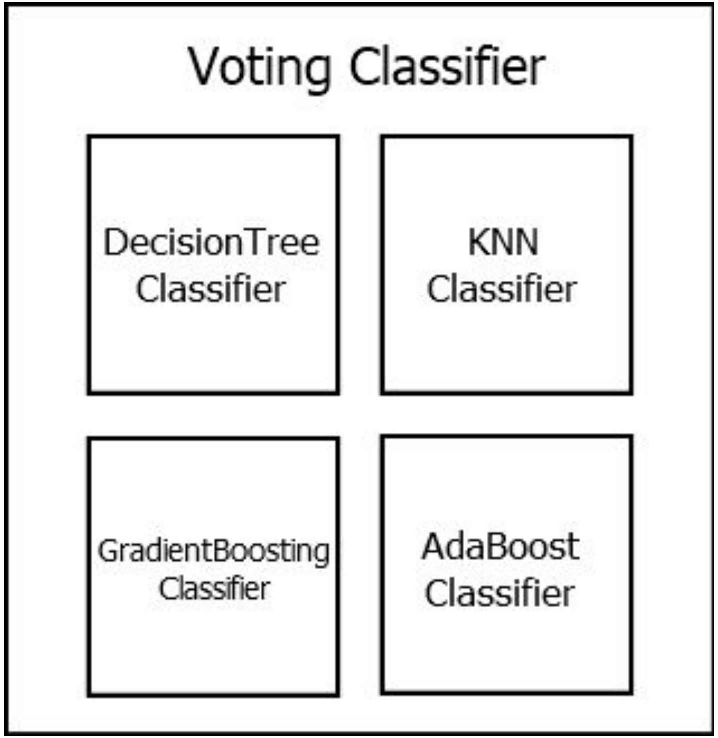
$$activities\_count\_mod = no\_of\_sales\_12M\_1 + no\_of\_Redemption\_12M\_1 \\ + (500 \times (no\_of\_sales\_12M\_10K + no\_of\_Redemption\_12M\_10K))$$

- An AUM value calculated by averaging the absolute values of asset classes and products. Since AUM features contained negative values, the negative values were treated as values that indicated the AUM magnitude, hence the absolute value was applied to all columns and then summed.

$$abs\_aum\_avg\_2 = \frac{\sum abs(aum\_p\_cols) + \sum abs(aum\_ac\_cols)}{2}$$

Appendix D - classification model design

- 4 base estimators contribute to overall classification
  - DecisionTreeClassifier
  - KNeighborsClassifier
  - GradientBoostingClassifier
  - AdaBoostClassifier
- Voting classifier - final model
  - Predicts classification

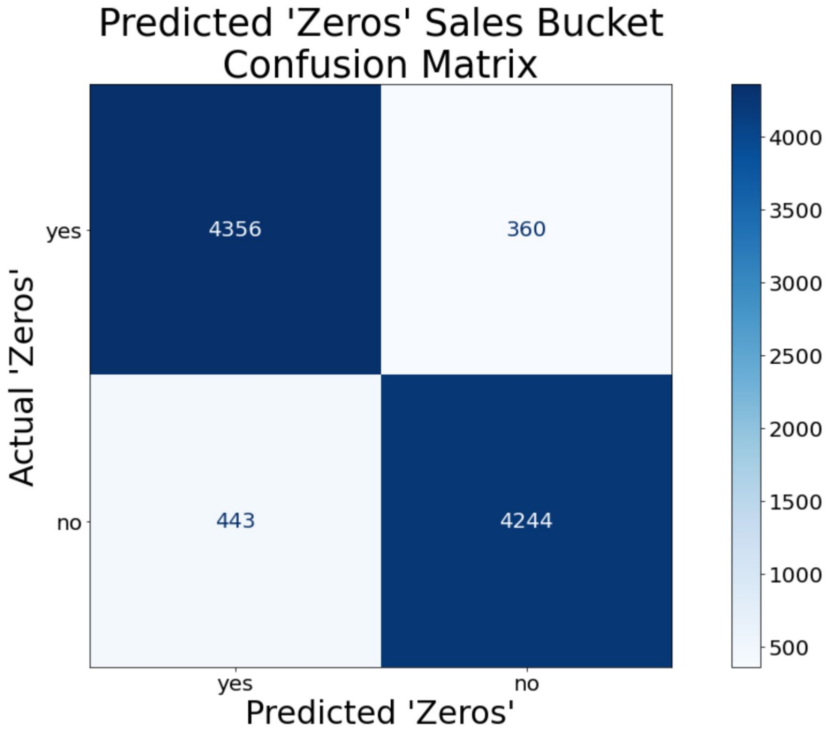


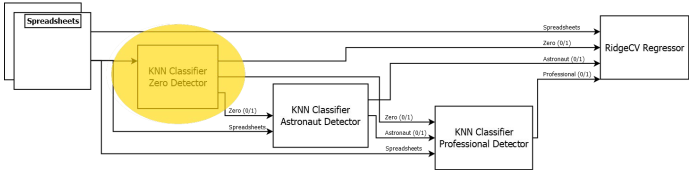
Appendix E - assessing classifier design

Classifier	Train Score (R^2)	Test Score (R^2)
DecisionTree	0.694	0.698
KNN	0.721	0.673
GradientBoost	0.765	0.755
AdaBoost	0.705	0.677
Voting (overall)	0.709	0.717

The ensemble voting classifier led to a robust design that avoided overfitting, and performed better than individual classifiers in terms of top loading the deciles 1 - 3 with a higher percentage of clients added funds.

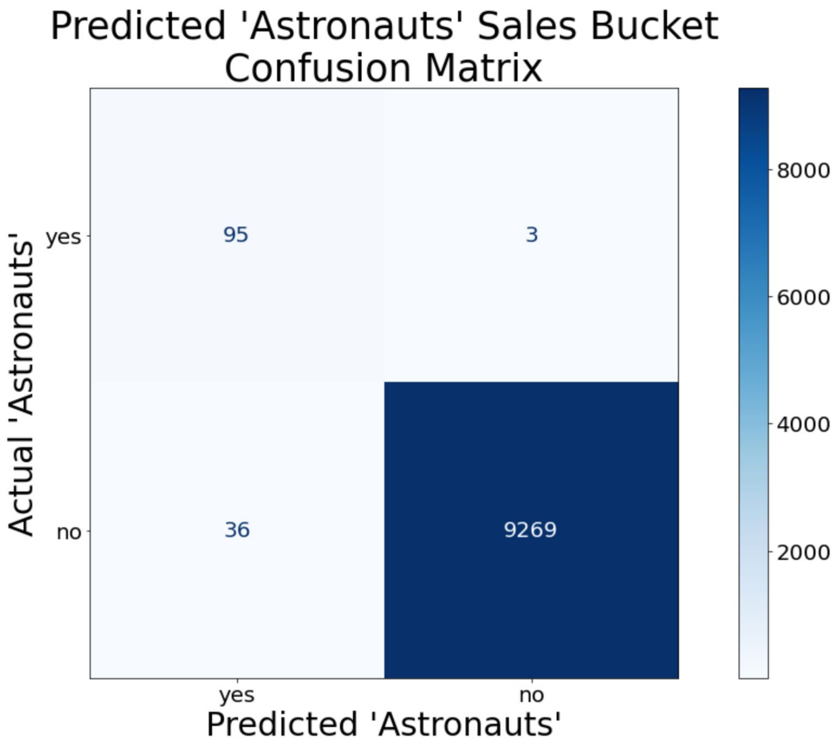
Appendix F - predicting zeros results

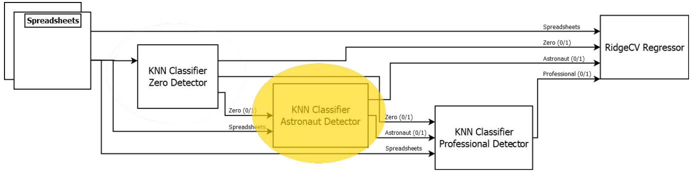




- Classifier scores R^2 (train, test)
  - (0.888, 0.876)
- Overall classification (including ‘curr_sales’ > 0 filtering) → 91.46% accuracy

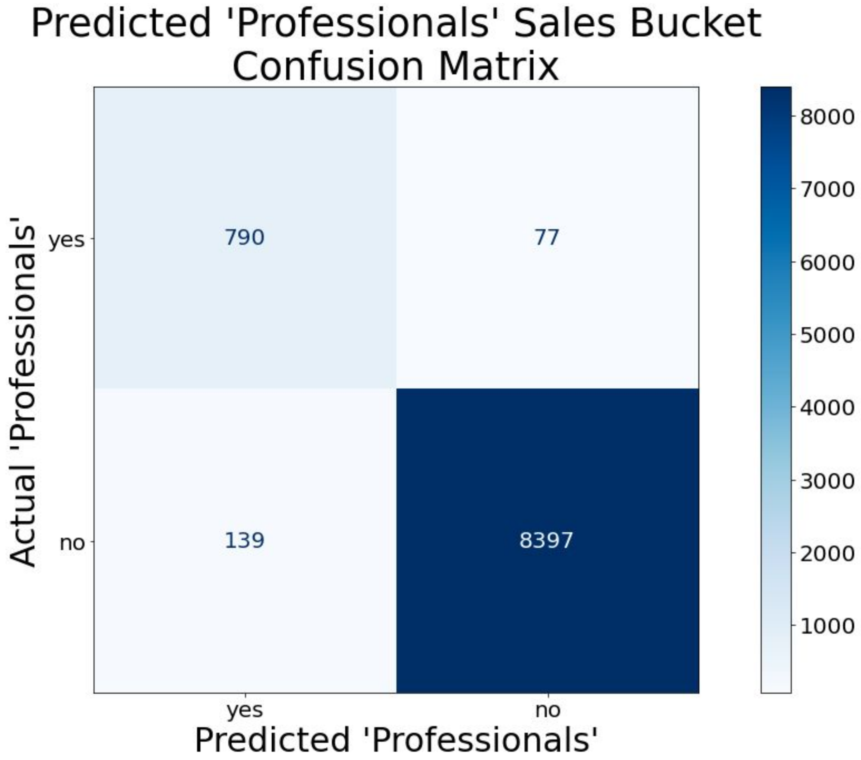
Appendix G - - predicting astronauts results



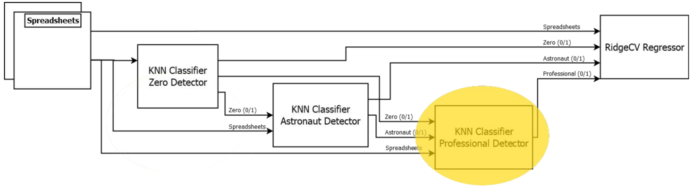


- Classifier scores R^2 (train, test)
  - (0.999, 0.986)

Appendix H - predicting professionals results



Predicted 'Professionals' Sales Bucket Confusion Matrix



- Classifier scores R^2 (train, test, baseline)
    - (0.993, 0.930, 0.901)

Appendix I - regression scores

Include spreadsheet prediction features	Include KNN classifier prediction features	Train Score (R^2)	Test Score (R^2)
1	0	0.586	0.561
0	1	0.705	0.677
1	1	0.729	0.701

- Including the KNN classifier predictions as features provides significant lift to model prediction
- The overall model (using both KNN classifier prediction features and other features directly from or derived from the spreadsheets) only performs slightly better than the KNN prediction features alone
    - All features test score 0.701, KNN classifiers features only test score 0.677
- Other features directly from or derived from spreadsheets include:
    - ‘abs_aum_avg_2’, ‘sales_total’, ‘sales_curr’, ‘No_of_fund_curr’, ‘activities_count_mod’

The End

No really, this is for real the end.