Shelter Animal Outcomes

Imelda Flores
Kristen Marenco

Introduction

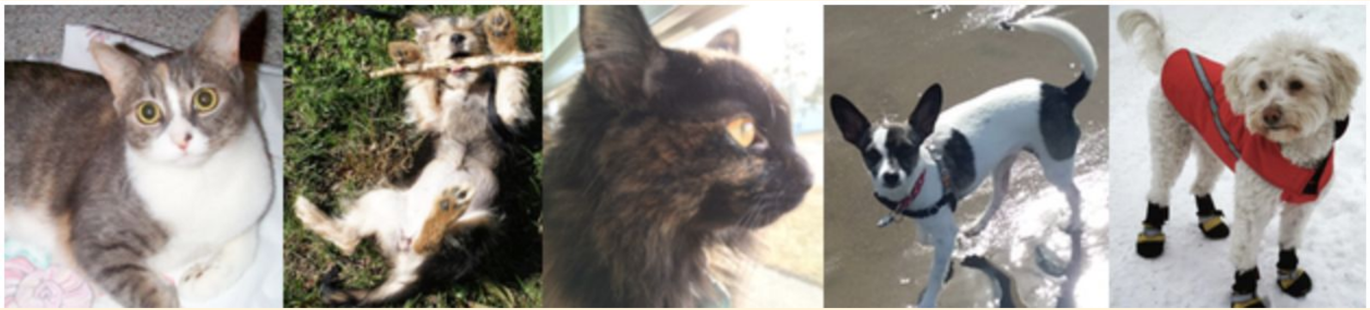
“Every year, approximately 7.6 million companion animals end up in US shelters. Many animals are given up as unwanted by their owners, while others are picked up after getting lost or taken out of cruelty situations. Many of theses animals find forever families to take them home, but just as many are not so lucky. 2.7 million dogs and cats are euthanized in the US every year “

Source: https://www.kaggle.com/c/shelter-animal-outcomes

Goal

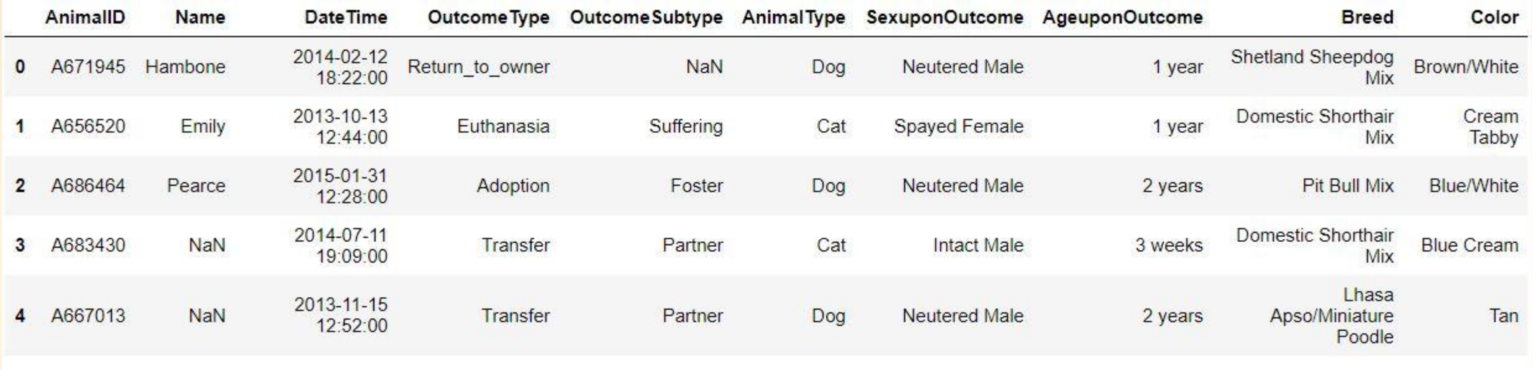
The goal of this project is to predict shelter animal outcomes such as:

- Adoption
- Transfer
- Euthanasia
- Died
- Return to owner



Dataset

1. Animal ID
2. Name
3. DateTime
4. OutcomeType
5. OutcomeSubtype
6. AnimalType
7. SexuponOutcome
8. AgeuponOutcome
9. Breed
10. Color



	AnimalID	Name	DateTime	OutcomeType	OutcomeSubtype	AnimalType	SexuponOutcome	AgeuponOutcome	Breed	Color
0	A671945	Hambone	2014-02-12 18:22:00	Return_to_owner	NaN	Dog	Neutered Male	1 year	Shetland Sheepdog Mix	Brown/White
1	A656520	Emily	2013-10-13 12:44:00	Euthanasia	Suffering	Cat	Spayed Female	1 year	Domestic Shorthair Mix	Cream Tabby
2	A686464	Pearce	2015-01-31 12:28:00	Adoption	Foster	Dog	Neutered Male	2 years	Pit Bull Mix	Blue/White
3	A683430	NaN	2014-07-11 19:09:00	Transfer	Partner	Cat	Intact Male	3 weeks	Domestic Shorthair Mix	Blue Cream
4	A667013	NaN	2013-11-15 12:52:00	Transfer	Partner	Dog	Neutered Male	2 years	Lhasa Apso/Miniature Poodle	Tan

Source: Austin Animal Center - 26,000 samples

Changes to Dataset

- DateTime:

  Hour, Day, Month, Year columns

- OneHotEncoding:

  Hour, AnimalType, SexuponOutcome, Breed, Color

- Units converted to days:

  AgeuponOutcome

- Boolean:

  Name: True if animal has name

  False if animal does not have name

New Dataset

1. Name
2. AgeinDaysUponOutcome
3. Day
4. Month
5. Year
6. Hour0
7. Hour1
8. Hour2
9. Hour3
10. Cat
11. Dog
12. IntactFemale
13. IntactMale
14. NeuteredMale
15. SpayedFemale
16. UnknownSex
17. Pit Bull
18. Chihuahua
19. Shepherd
20. Retriever
21. Terrier
22. DomesticShorthair
23. DomesticMediumHair
24. DomesticLonghair
25. Siamese
26. Other Breed
27. Black
28. Brown
29. White
30. Tan
31. Blue
32. Tabby
33. Other Color

Label: OutcomeType

Predictive Models

Imelda Flores

1. Support Vector Machine (SVM)
2. GridSearchCV
3. Gradient Boosting Classifier

Kristen Marenco

1. Random Forest
2. XGBoost

SVM

- **Why SVM?**

  SVM was chosen because it does some extremely complex data transformations, since it converts not separable problem to separable problem, these functions are called kernels.

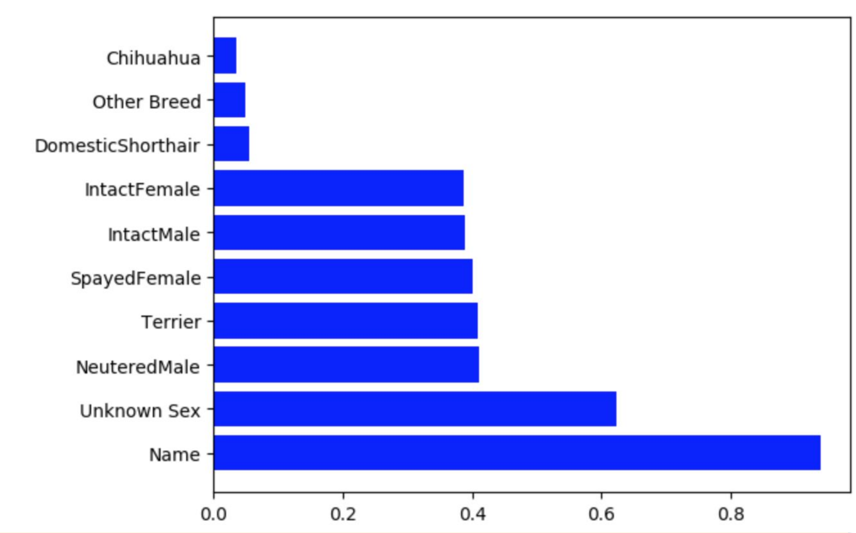
  Accuracy: 64.03%

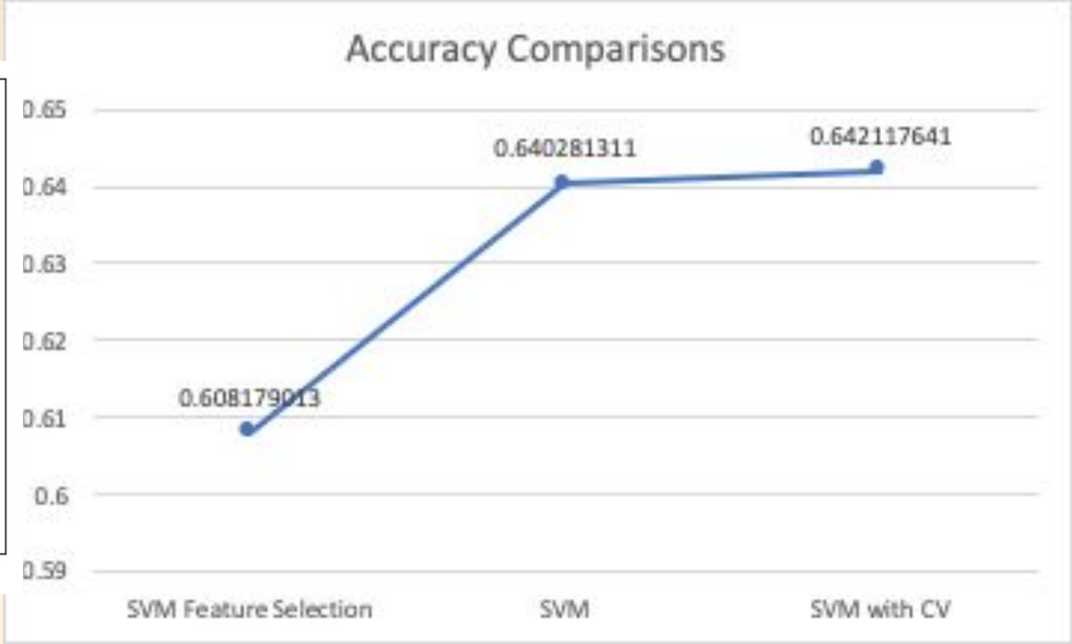
- **SVM with Cross Validation:**

  Accuracy: 64.20%

SVM Continued

- **Feature Selection:**
  - features: ['Name', 'Unknown Sex', 'NeuteredMale', 'Terrier', 'SpayedFemale']
  - Accuracy: 60.82%



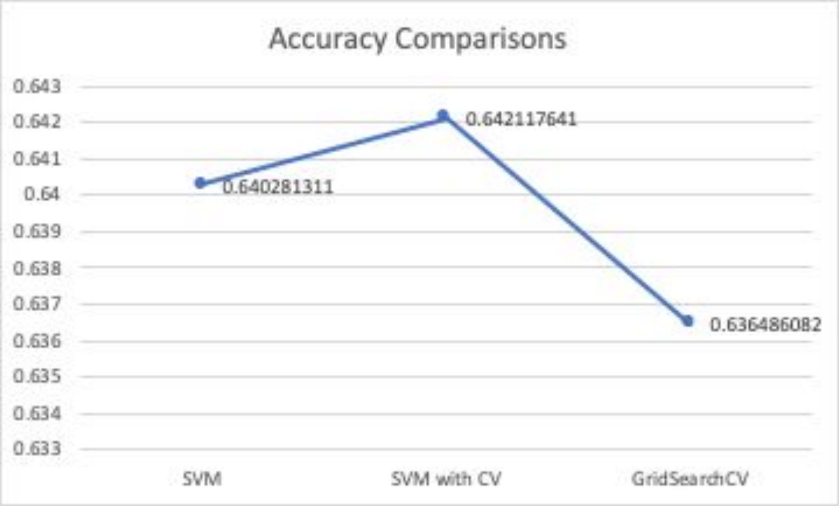


SVM and GridSearchCV

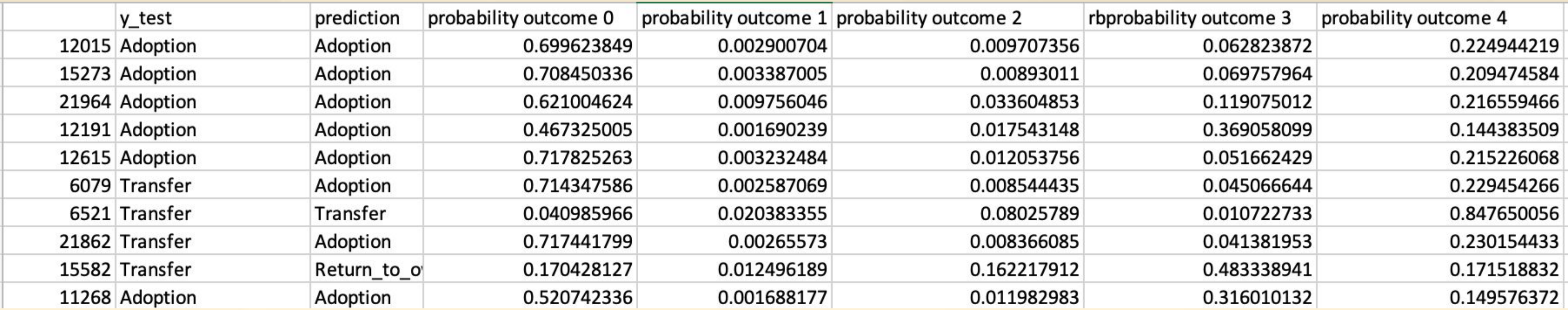
- **Why GridSearch?**

GridSearch will try the possible combinations with the given parameters and return the highest accuracy found. GridSearch can be slow since it has to execute a large amount of combinations.

SVM and GridSearch accuracy 63.645%



SVM Continued



	y_test	prediction	probability outcome 0	probability outcome 1	probability outcome 2	rbprobability outcome 3	probability outcome 4
12015	Adoption	Adoption	0.699623849	0.002900704	0.009707356	0.062823872	0.224944219
15273	Adoption	Adoption	0.708450336	0.003387005	0.00893011	0.069757964	0.209474584
21964	Adoption	Adoption	0.621004624	0.009756046	0.033604853	0.119075012	0.216559466
12191	Adoption	Adoption	0.467325005	0.001690239	0.017543148	0.369058099	0.144383509
12615	Adoption	Adoption	0.717825263	0.003232484	0.012053756	0.051662429	0.215226068
6079	Transfer	Adoption	0.714347586	0.002587069	0.008544435	0.045066644	0.229454266
6521	Transfer	Transfer	0.040985966	0.020383355	0.08025789	0.010722733	0.847650056
21862	Transfer	Adoption	0.717441799	0.00265573	0.008366085	0.041381953	0.230154433
15582	Transfer	Return_to_o	0.170428127	0.012496189	0.162217912	0.483338941	0.171518832
11268	Adoption	Adoption	0.520742336	0.001688177	0.011982983	0.316010132	0.149576372

Gradient Boosting Classifier

- **Why?**

  The classifier calculates error and update the weights to minimize the error. A tree is added to reduce the loss and recalculated after it’s added. When the loss is at a level that no longer improves with the dataset then a fixed number of trees is added or the training stops.

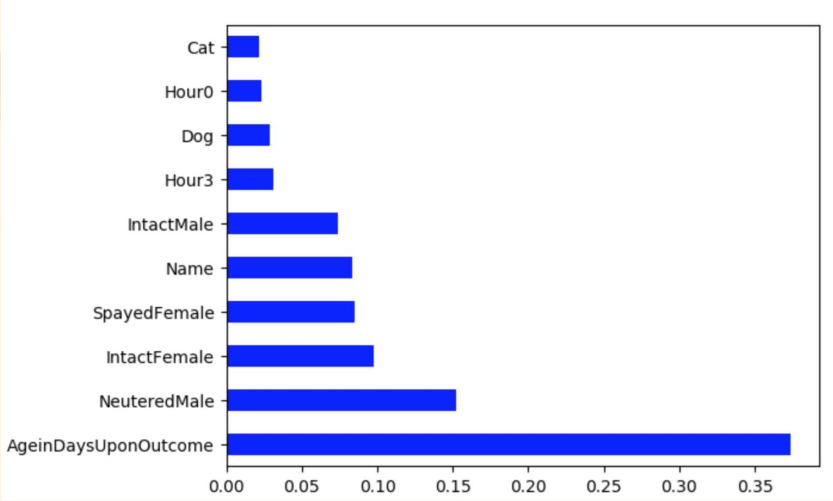
  Accuracy: 63.92%

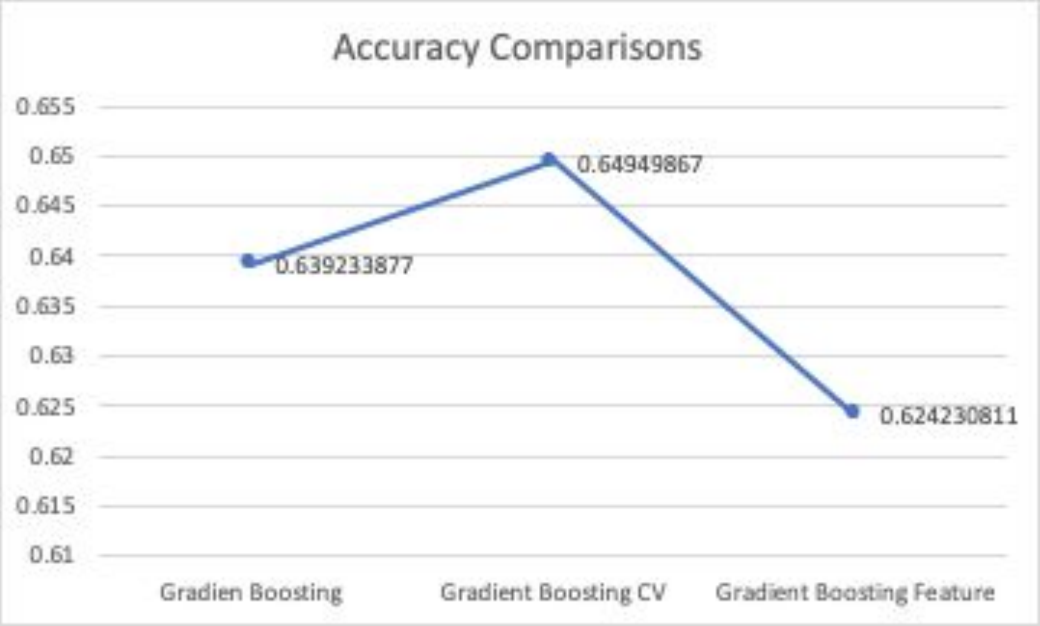
- **Gradient Boosting with Cross Validation:**

  Accuracy: 64.95%

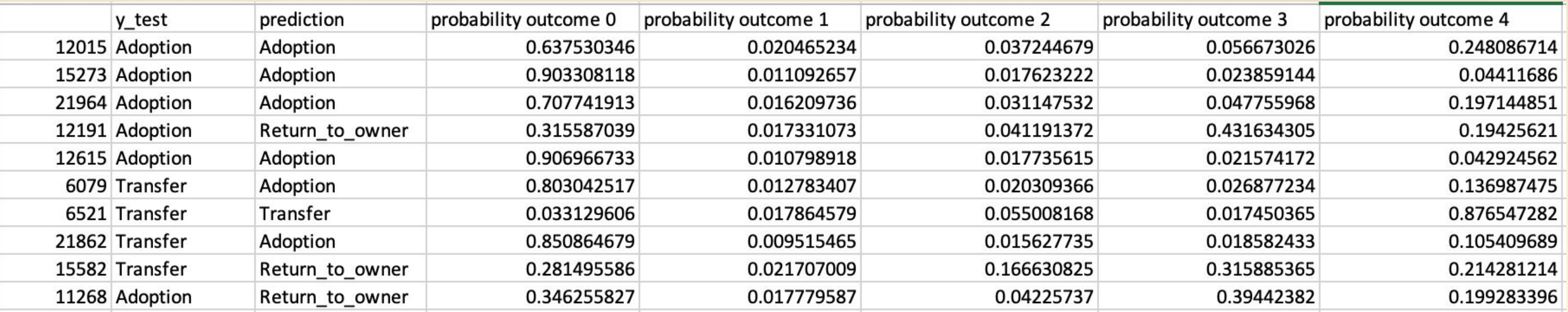
Gradient Boosting Classifier Continued

- **Feature Selection:**
  - features: ['AgeinDaysUponOutcome', 'NeuteredMale', 'IntactFemale', 'SpayedFemale', 'Name']
  - Accuracy: 62.42%





Gradient Boosting Classifier Continued



	y_test	prediction	probability outcome 0	probability outcome 1	probability outcome 2	probability outcome 3	probability outcome 4
12015	Adoption	Adoption	0.637530346	0.020465234	0.037244679	0.056673026	0.248086714
15273	Adoption	Adoption	0.903308118	0.011092657	0.017623222	0.023859144	0.04411686
21964	Adoption	Adoption	0.707741913	0.016209736	0.031147532	0.047755968	0.197144851
12191	Adoption	Return_to_owner	0.315587039	0.017331073	0.041191372	0.431634305	0.19425621
12615	Adoption	Adoption	0.906966733	0.010798918	0.017735615	0.021574172	0.042924562
6079	Transfer	Adoption	0.803042517	0.012783407	0.020309366	0.026877234	0.136987475
6521	Transfer	Transfer	0.033129606	0.017864579	0.055008168	0.017450365	0.876547282
21862	Transfer	Adoption	0.850864679	0.009515465	0.015627735	0.018582433	0.105409689
15582	Transfer	Return_to_owner	0.281495586	0.021707009	0.166630825	0.315885365	0.214281214
11268	Adoption	Return_to_owner	0.346255827	0.017779587	0.04225737	0.39442382	0.199283396

Random Forest

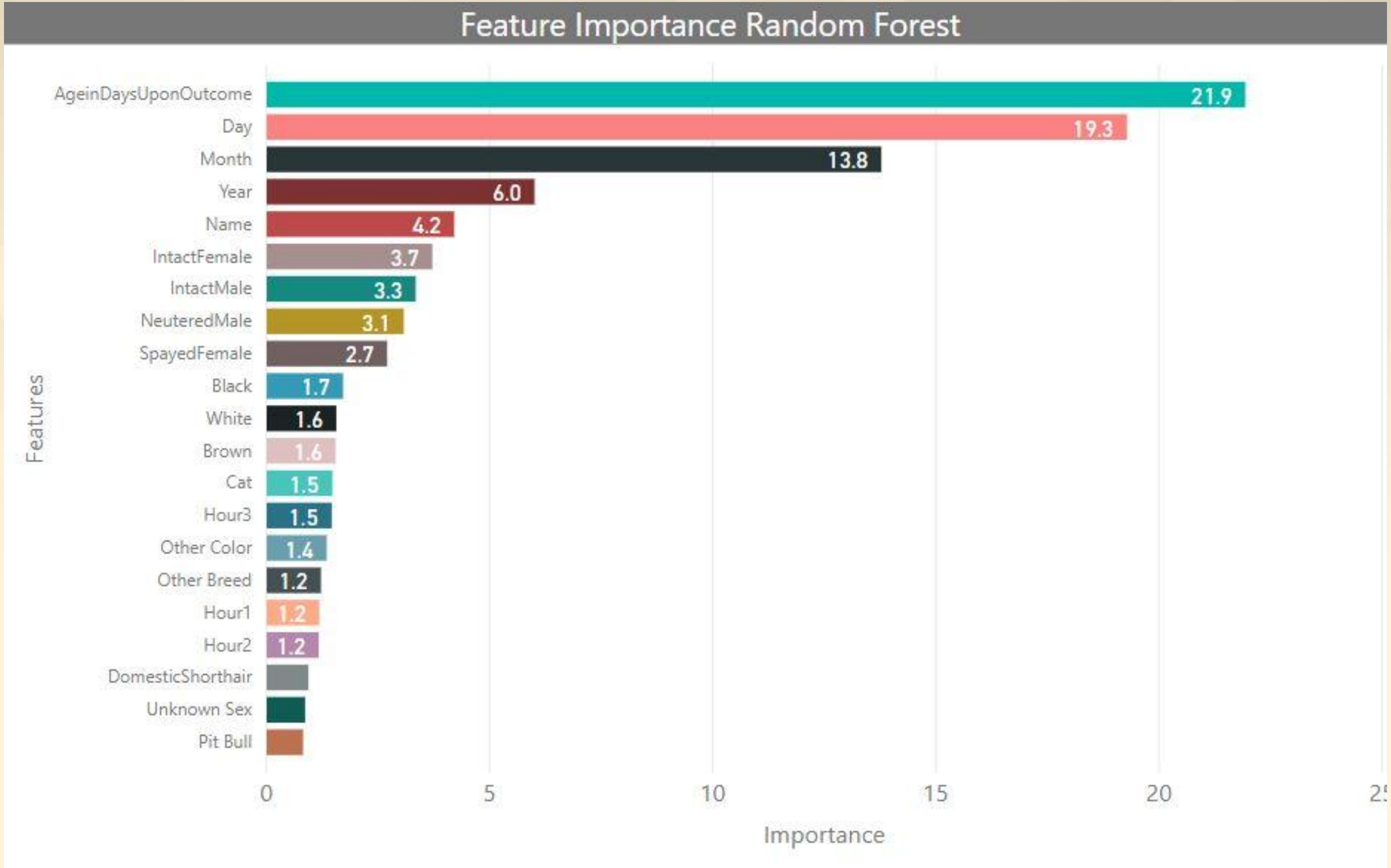
Why Choose Random Forest?

Generally produces a good predictive model, avoids overfitting, uses bagging, and a diverse set of decision trees.

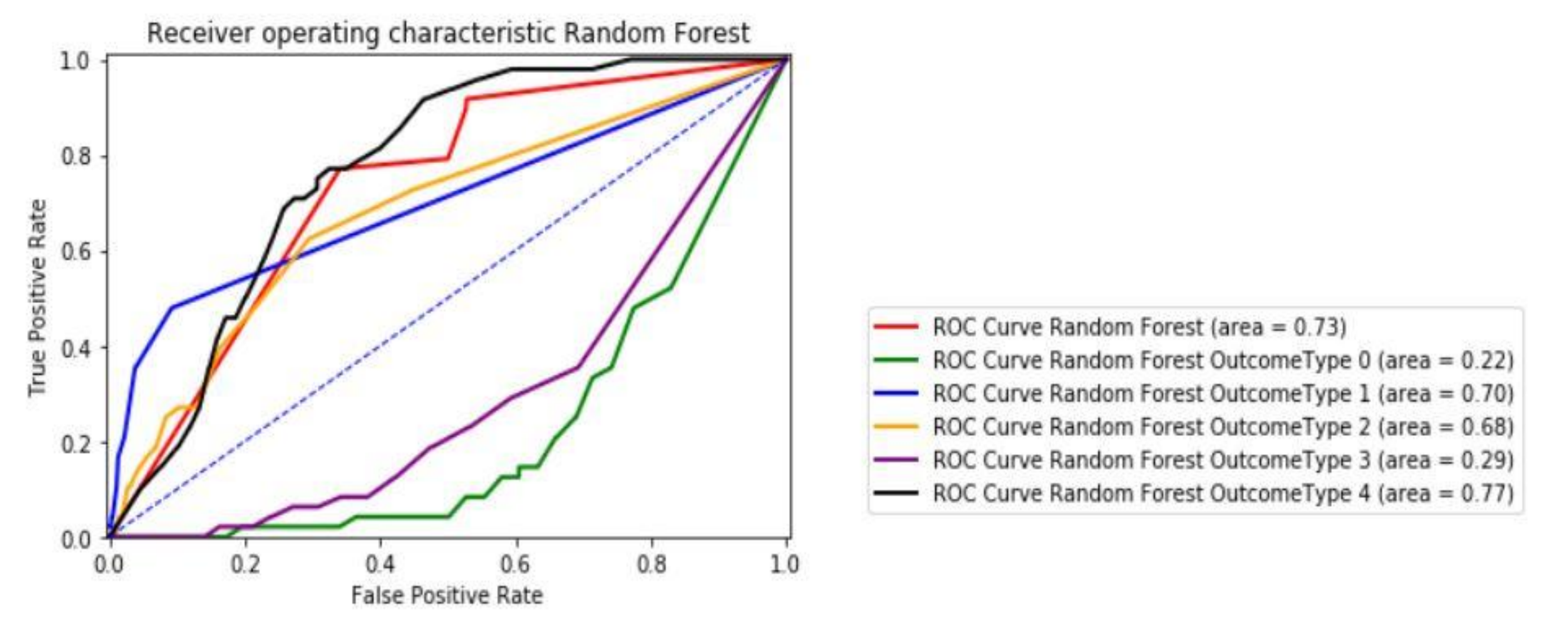
10- Fold Cross Validation Accuracy of Random Forest:

62.50%

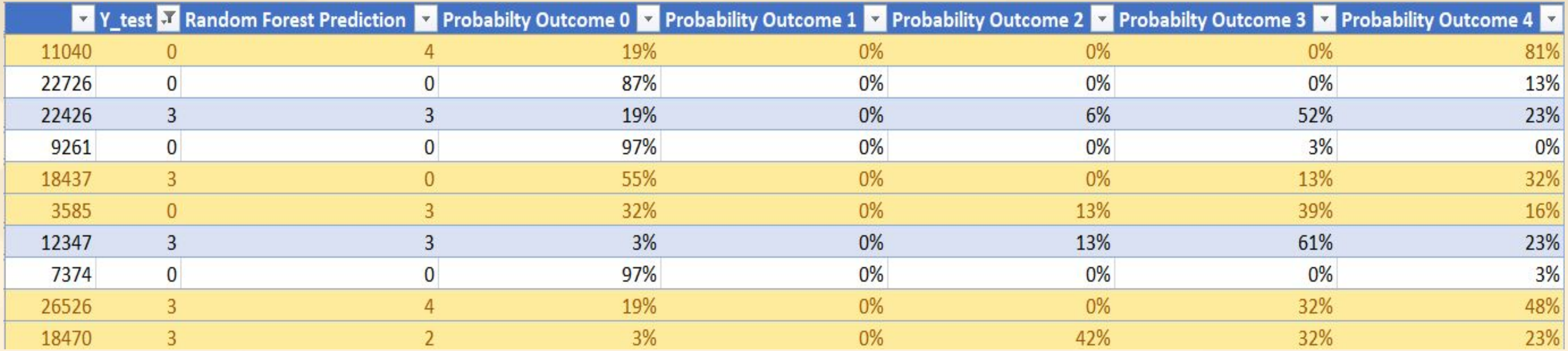
Random Forest Continued



Random Forest Continued



Random Forest Continued



	Y_test	Random Forest Prediction	Probabilty Outcome 0	Probability Outcome 1	Probability Outcome 2	Probabilty Outcome 3	Probability Outcome 4
11040	0	4	19%	0%	0%	0%	81%
22726	0	0	87%	0%	0%	0%	13%
22426	3	3	19%	0%	6%	52%	23%
9261	0	0	97%	0%	0%	3%	0%
18437	3	0	55%	0%	0%	13%	32%
3585	0	3	32%	0%	13%	39%	16%
12347	3	3	3%	0%	13%	61%	23%
7374	0	0	97%	0%	0%	0%	3%
26526	3	4	19%	0%	0%	32%	48%
18470	3	2	3%	0%	42%	32%	23%

XGBoost

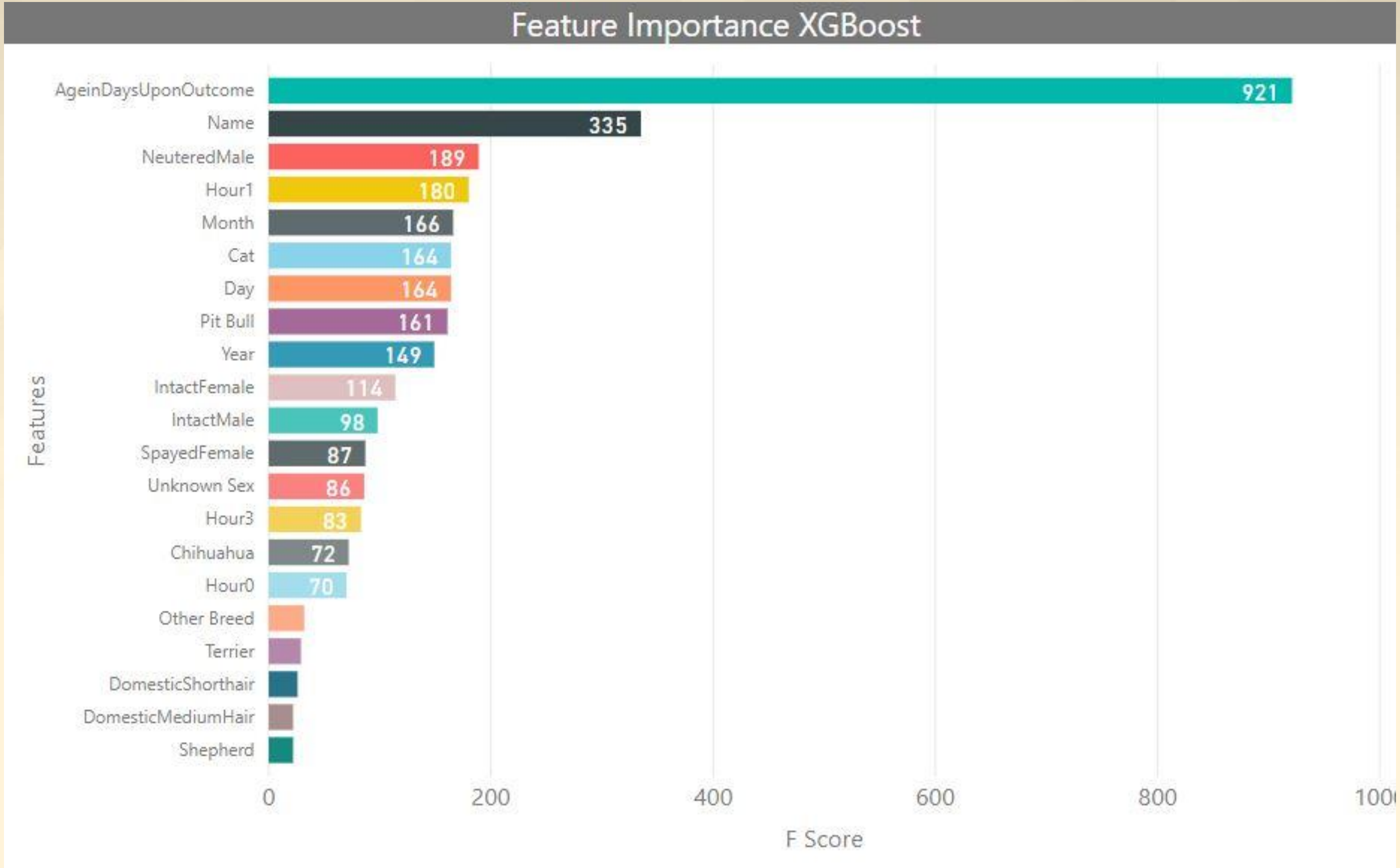
Why Choose XGBoost?

Boosts a set of decision trees considered to be weak learners into strong learners, using continued training and voting.

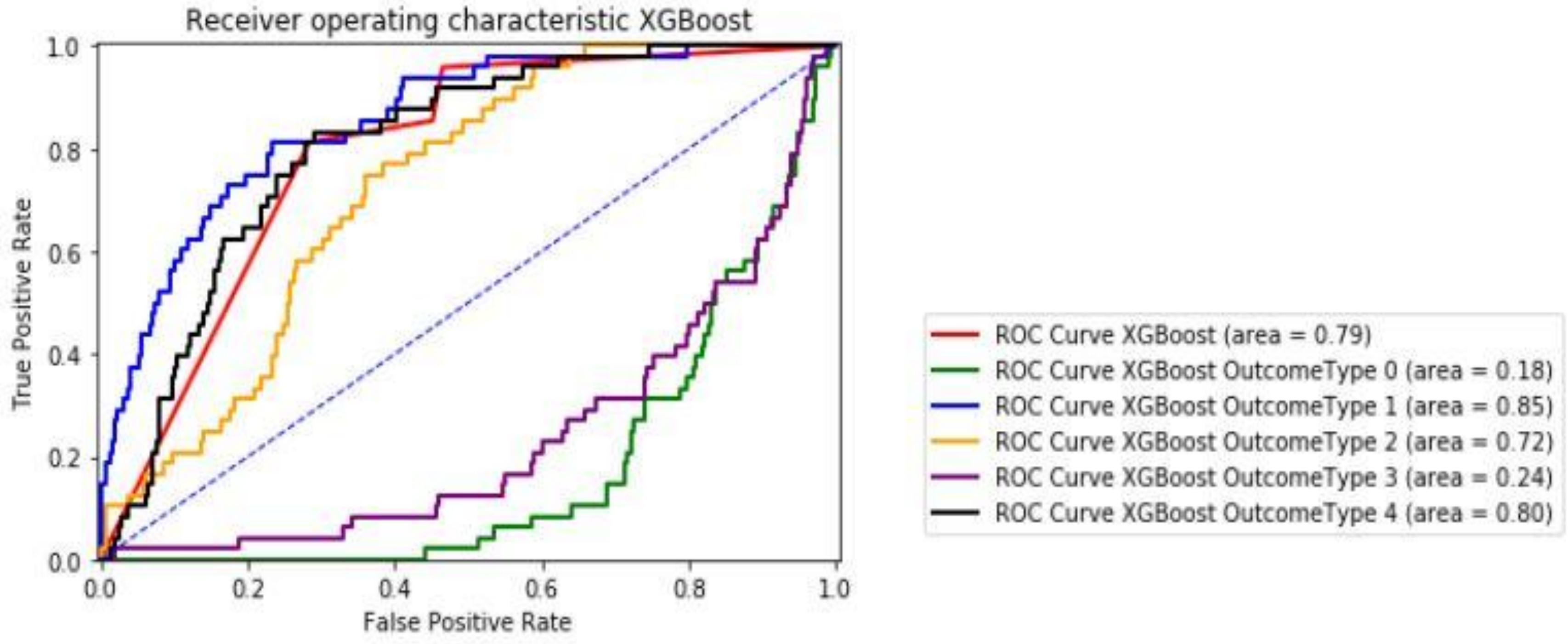
10- Fold Cross Validation Accuracy of XGBoost:

65.75%

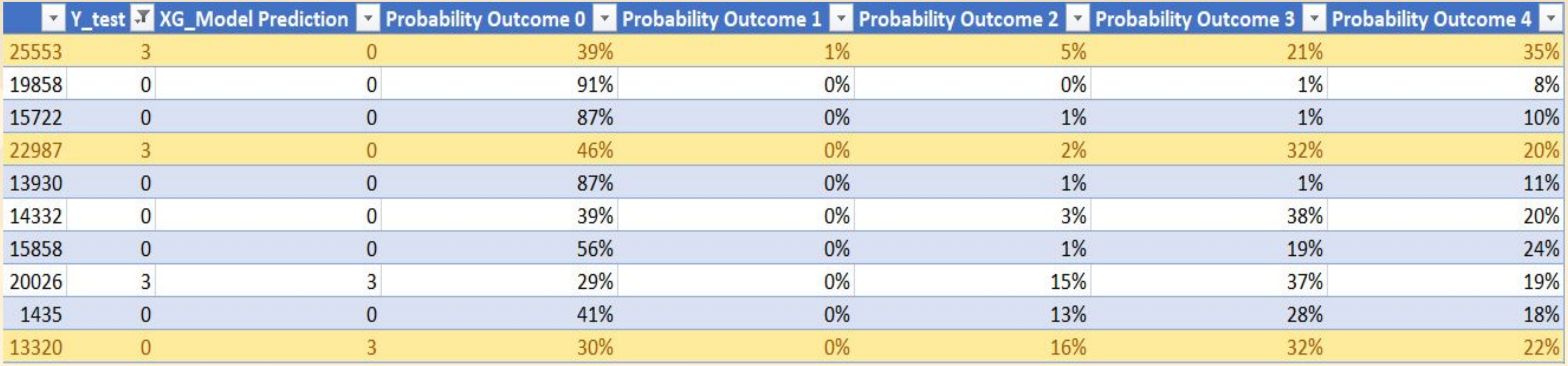
XGBoost Continued



XGBoost Continued

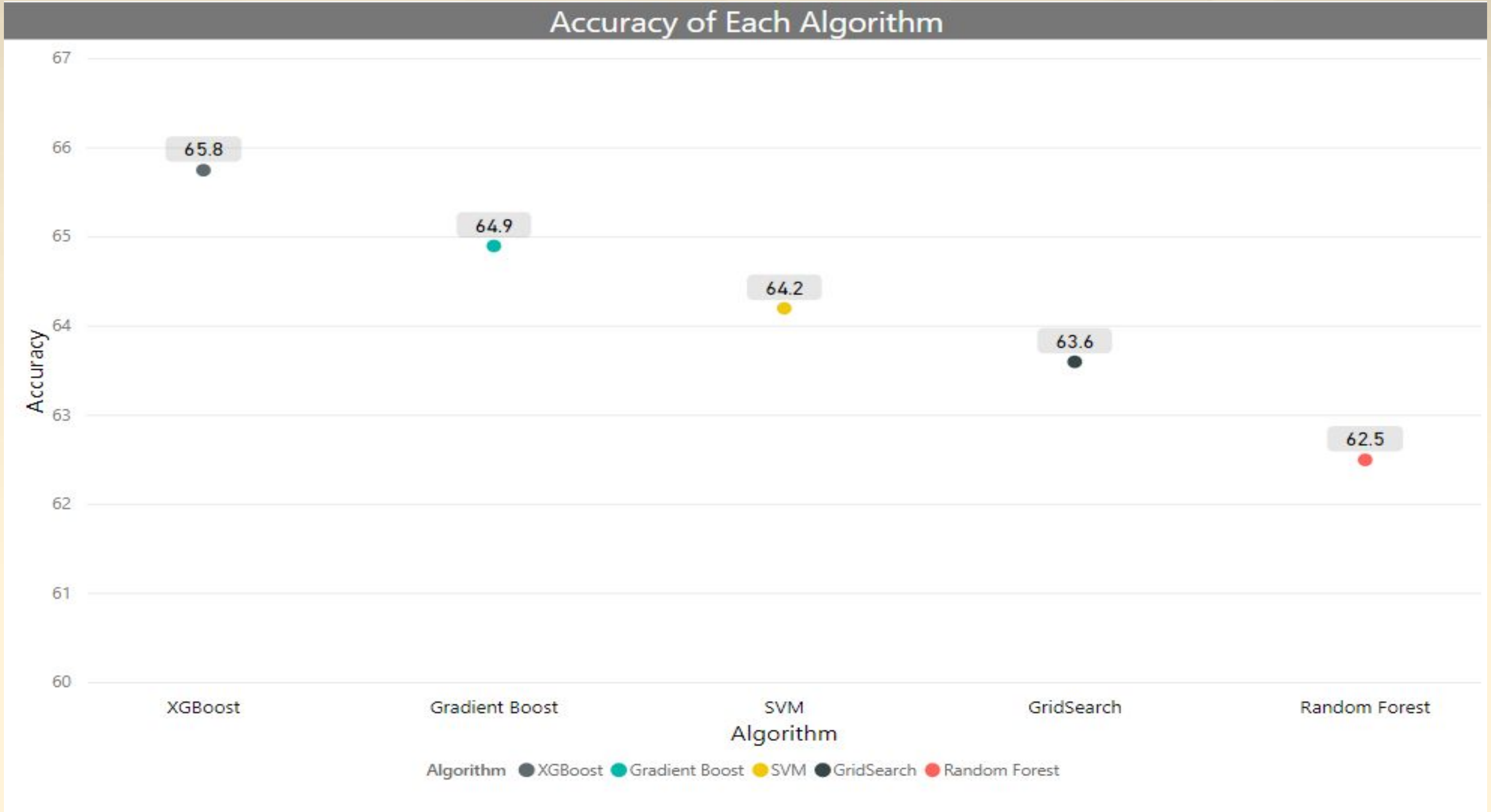


XGBoost Continued



	Y_test	XG_Model Prediction	Probability Outcome 0	Probability Outcome 1	Probability Outcome 2	Probability Outcome 3	Probability Outcome 4
25553	3	0	39%	1%	5%	21%	35%
19858	0	0	91%	0%	0%	1%	8%
15722	0	0	87%	0%	1%	1%	10%
22987	3	0	46%	0%	2%	32%	20%
13930	0	0	87%	0%	1%	1%	11%
14332	0	0	39%	0%	3%	38%	20%
15858	0	0	56%	0%	1%	19%	24%
20026	3	3	29%	0%	15%	37%	19%
1435	0	0	41%	0%	13%	28%	18%
13320	0	3	30%	0%	16%	32%	22%

Team Results



Conclusion

Why was the accuracy so low?

- Datetime was considered data leakage, however Kaggle still advised to to include the column in training.

- The Color column was jumbled up with values and was difficult to include all values.

- Outcome Type 1 only made up 0.7% of the dataset, not providing enough samples to classify the type correctly.

- Too many overlapping features, there were not enough distinguishable features in the dataset for models to distinguish between outcomes.

Q & A