GROWING (FOR) CONCERN

INTRODUCTION

Data Used:

- SAVSNET Tumour Database

Tools Utilized:

- Python Pandas & Numpy for data manipulation
- Python Seaborn and Matplotlib for visualizations
- Python Scikit-learn for modeling and ML
- Python Seaborn and Matplotlib for visualizations

METHODOLOGY

- **Data Collection and Cleaning**
  - **Tumor data sourced from relatively new database of pet cancer cases provided by SAVSNET: Small Animal Veterinary Surveillance Network**
- **Feature Engineering**
  - **Segmented categorical data to try and determine indicators of multiple tumor count and potential high/low risk features**
- **Modeling and Selection**
  - **Trained an ensemble voting classifier to determine likelihood of multiple lesions**

PROCESS

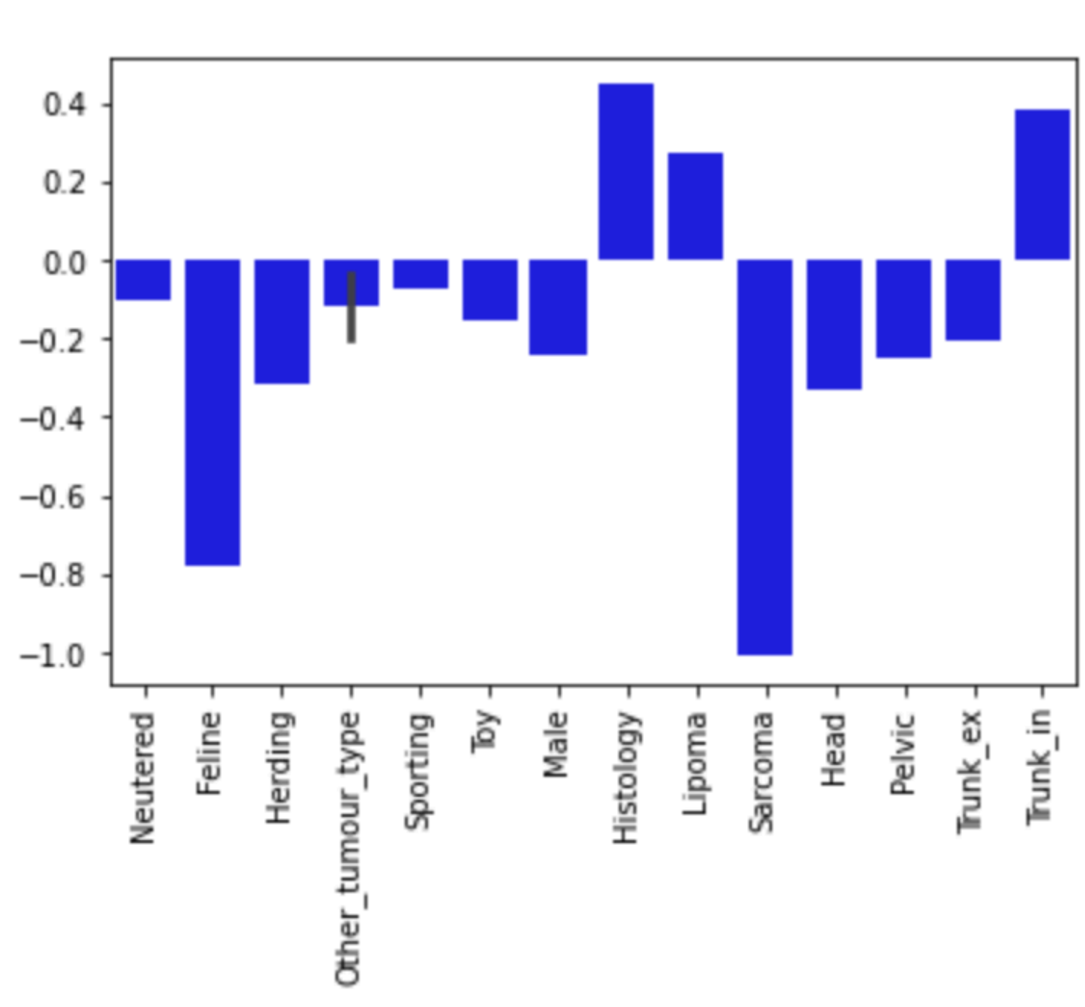
ORIGINAL DATA

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 109895 entries, 0 to 109894
Data columns (total 10 columns):
 #   Column                  Non-Null Count   Dtype
---  ------                  --------------   -----
 0   Species                 109895 non-null  object
 1   Breed                   109895 non-null  object
 2   Gender                  109895 non-null  object
 3   Histo_Cyto              109895 non-null  object
 4   Tumours_in_the_report   109895 non-null  int64
 5   Primary_tumour          109895 non-null  object
 6   Grade_2_tier            13539 non-null   object
 7   Grade_3_tier            9660 non-null    object
 8   Location                108793 non-null  object
 9   Neutered                109895 non-null  int64
```

Original dataset included various traits about prior cases including physical traits of the animal as well as how and where the tumor was found.

A vast majority of these needed further segmentation to become viable features, and a number were immediately removed (e.g. anonymous practice ID)

HIGHEST/LOWEST RISK FEATURES



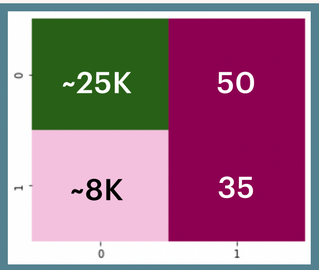
- These features provided the final baseline for modeling and learning
- Notable features which did not provide a signal either way:
    - Non-sporting Breed Group
    - Lymphoma type Tumour
    - The Lymph System Location

RESULTS

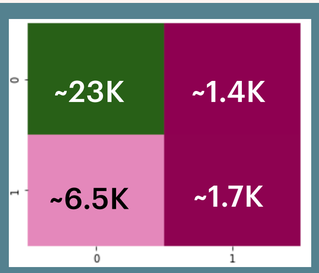
DATA SELECTION & MODEL METRICS

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 109895 entries, 0 to 109894
Data columns (total 15 columns):
 #    Column               Non-Null Count   Dtype
---   ------               --------------   -----
 0    Neutered             109895 non-null  int64
 1    Feline               109895 non-null  uint8
 2    Herding              109895 non-null  uint8
 3    Other_tumour_type    109895 non-null  uint8
 4    Sporting             109895 non-null  uint8
 5    Toy                  109895 non-null  uint8
 6    Male                 109895 non-null  uint8
 7    Histology            109895 non-null  uint8
 8    Lipoma               109895 non-null  uint8
 9    Other_tumour_type    109895 non-null  uint8
 10   Sarcoma              109895 non-null  uint8
 11   Head                 109895 non-null  uint8
 12   Pelvic               109895 non-null  uint8
 13   Trunk_ex             109895 non-null  uint8
 14   Trunk_in             109895 non-null  uint8
```

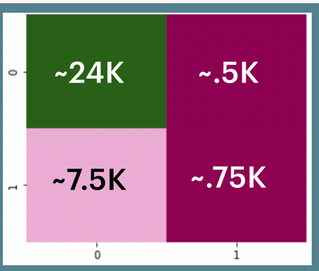
- Logistic Regression
- Accuracy score of .749



- Decision Tree
- Accuracy score of .759



- Random Forest
- Accuracy score of .749

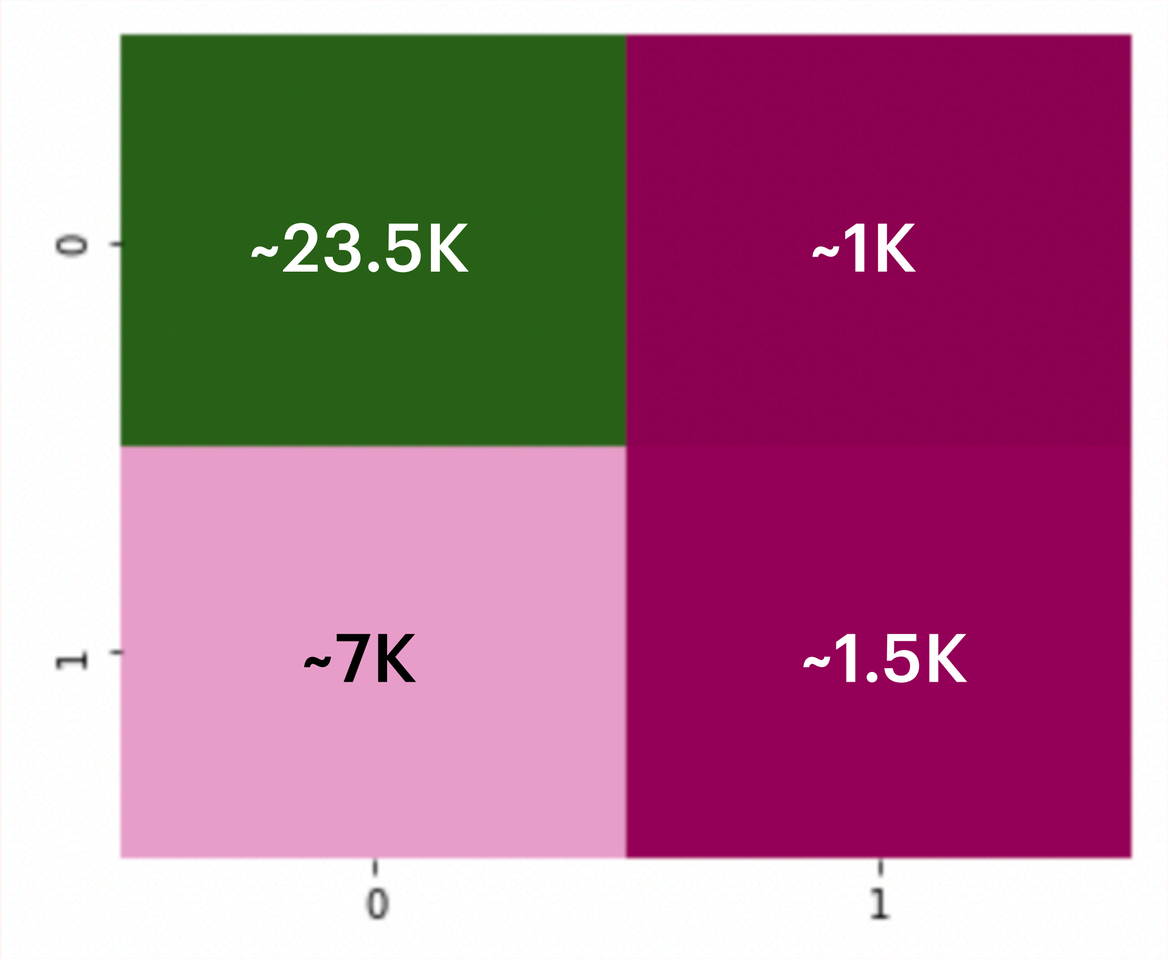


FINAL MODEL

Final Model

- Voting Classifier Ensemble
  - Decision tree weighted at 2 captures more true positives, but struggles to classify negatives accurately
  - Logistic Regression weighted at 1.4 helps compensate for negative difficulties in Decision Tree

Accuracy of .7603 with highest confidence on both Positive and Negative predictions.



FURTHER ACTION AND NOTEABLES

- **Low Risk Features:**
  - **Felines significantly more likely to only have one lesion**
  - **Sarcoma almost always only presents in one growth**
  - **Herding Group dogs less likely to have multiple presentations**

- **High Risk Features:**
  - **Multi-growth Cancers are more likely to effect organs and connective tissue than external**
  - **Lipoma is most likely to present in multiples**
  - **Age not provided in data, but definitely would affect.**

Further data acquisition concerning age could temper the results shown here, as increased age in all animals is linked to increase chance of cancer.

Further work with XGBoost would also likely assist in refining negative accuracy overall, and reduce false negatives