Flip00 Project Final Presentation

Shuxia Lin
SouthEast University

2019-10-27

Problem Statement
Exploratory Data Analysis
Methods
Forecast Results
Conclusion
Thanks for attending and welcome for questions

Overview

Problem Statement
- Problem Definition
- Data Set

Exploratory Data Analysis
- Data Visualization
- Data Engineering

Methods

Forecast Results
- Evaluation Methods
- Forecast Results

Conclusion

Thanks for attending and welcome for questions

Problem Statement
Problem Definition
Data Set
Exploratory Data Analysis
Methods
Forecast Results
Conclusion
Thanks for attending and welcome for questions

Problem Statement

Problem Definition

Problem Statement
Problem Definition
Data Set
Exploratory Data Analysis
Methods
Forecast Results
Conclusion
Thanks for attending and welcome for questions

Defn After a month of making scientific observations and taking careful measurements, can determined that 900 ghouls, ghosts, and goblins. The raw dataset contains train set with 371 samples and 529 unlabeled samples as test set. Through the train data, find the relationship between the attributes and species, and then identify the ghastly creatures in test data.

Data Set

Problem Statement
Problem Definition
Data Set
Exploratory Data Analysis
Methods
Forecast Results
Conclusion
Thanks for attending and welcome for questions

Defn There are 4 numerical variables and 1 categorical, and no missing values. Numerical columns are either normalized or show a percentage, so no need to scale them.

- Data List

  id id of the creature
  bone_length average length of bone in the creature, normalized between *0* and *1*
  rotting_flesh percentage of rotting flesh in the creature
  hair_length average hair length, normalized between *0* and *1*
  has_soul percentage of soul in the creature
  Color dominant color of the creature: *white*,*black*,*clear*, *blue*,*green*,*blood*
  type target variable: *Ghost*, *Goblin*, and *Ghoul*

- Train Data and Test Data

  Divide the raw train set into train data and test data, and the ratio is 8:2.

Problem Statement
Exploratory Data Analysis
Data Visualization
Data Engineering
Methods
Forecast Results
Conclusion
Thanks for attending and welcome for questions

Exploratory Data Analysis

Problem Statement
Exploratory Data Analysis
Data Visualization
Data Engineering
Methods
Forecast Results
Conclusion
Thanks for attending and welcome for questions

Data Visualization

Exp Use EDA to plot the distribution of the data, can observate the data intuitively and find the relation between the attribute values.

- Figures
  - Histogrm
  - Boxplot
  - Pairplot
  - Correllogram

Data Visualization

Problem Statement

Exploratory Data Analysis

Data Visualization

Data Engineering

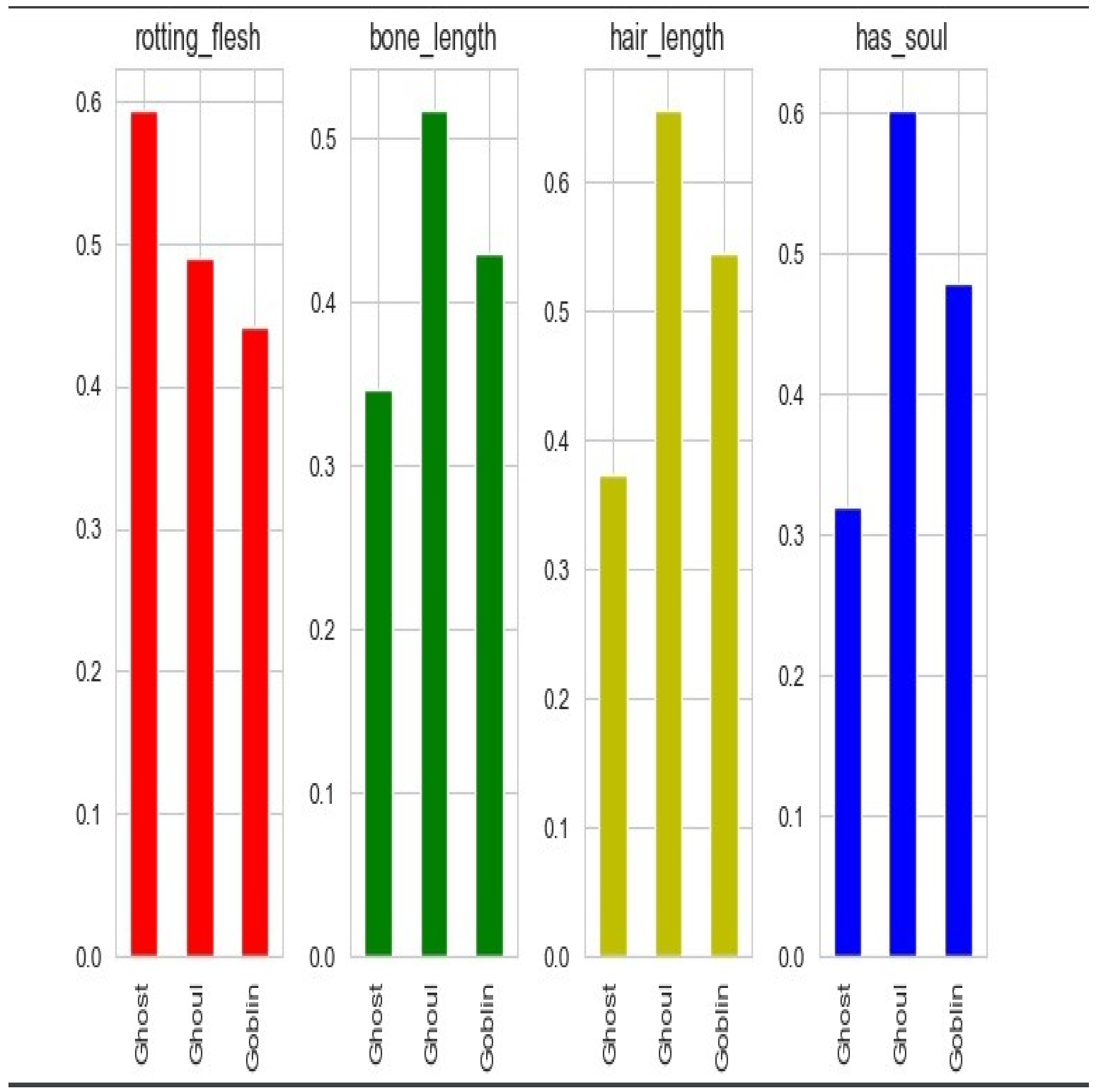
Methods

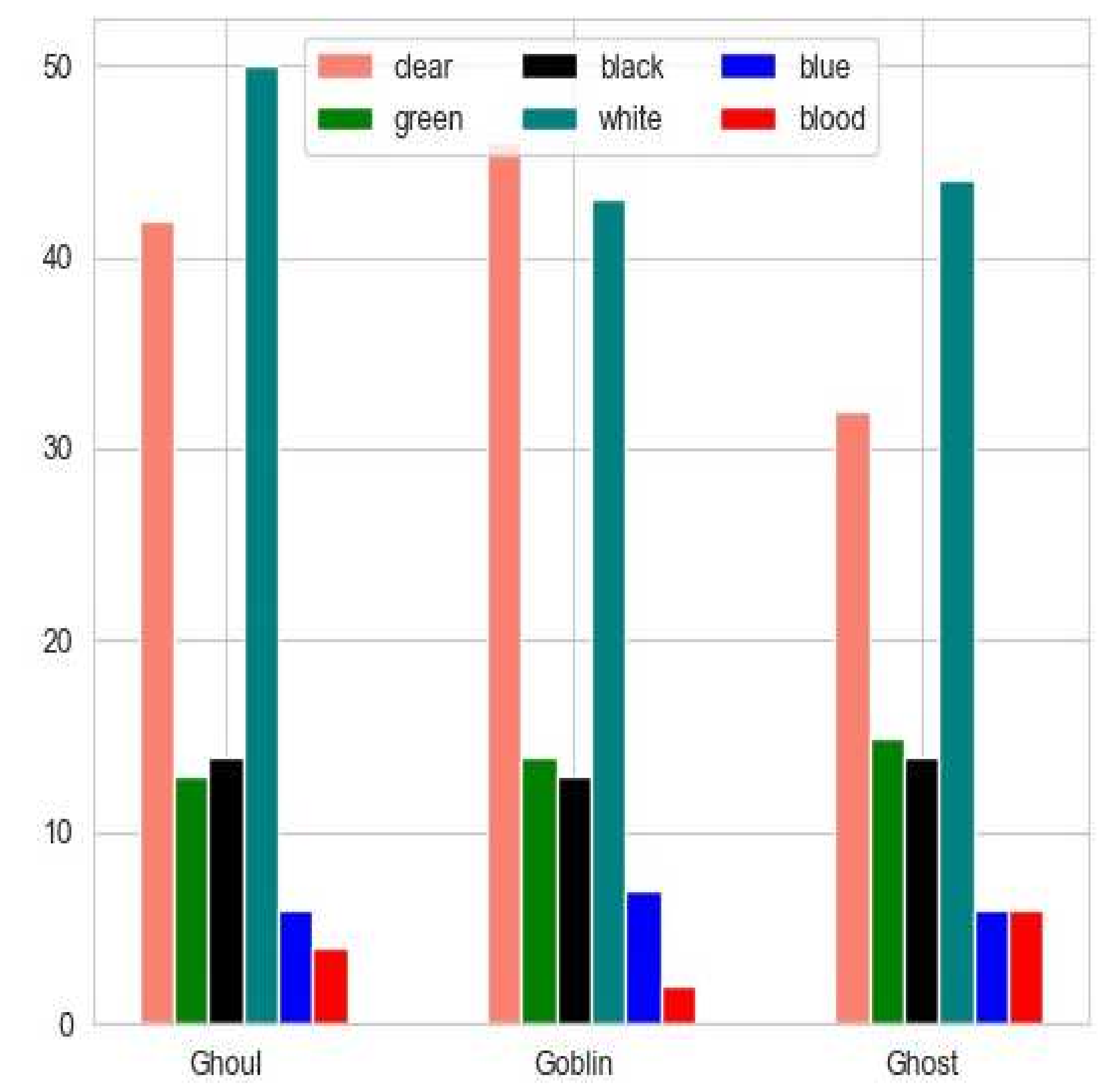
Forecast Results

Conclusion

Thanks for attending and welcome for questions

Exp It seems that all numerical features may be useful, but many colors are evenly distributes among the monsters, which means they maybe have little effect on classification.



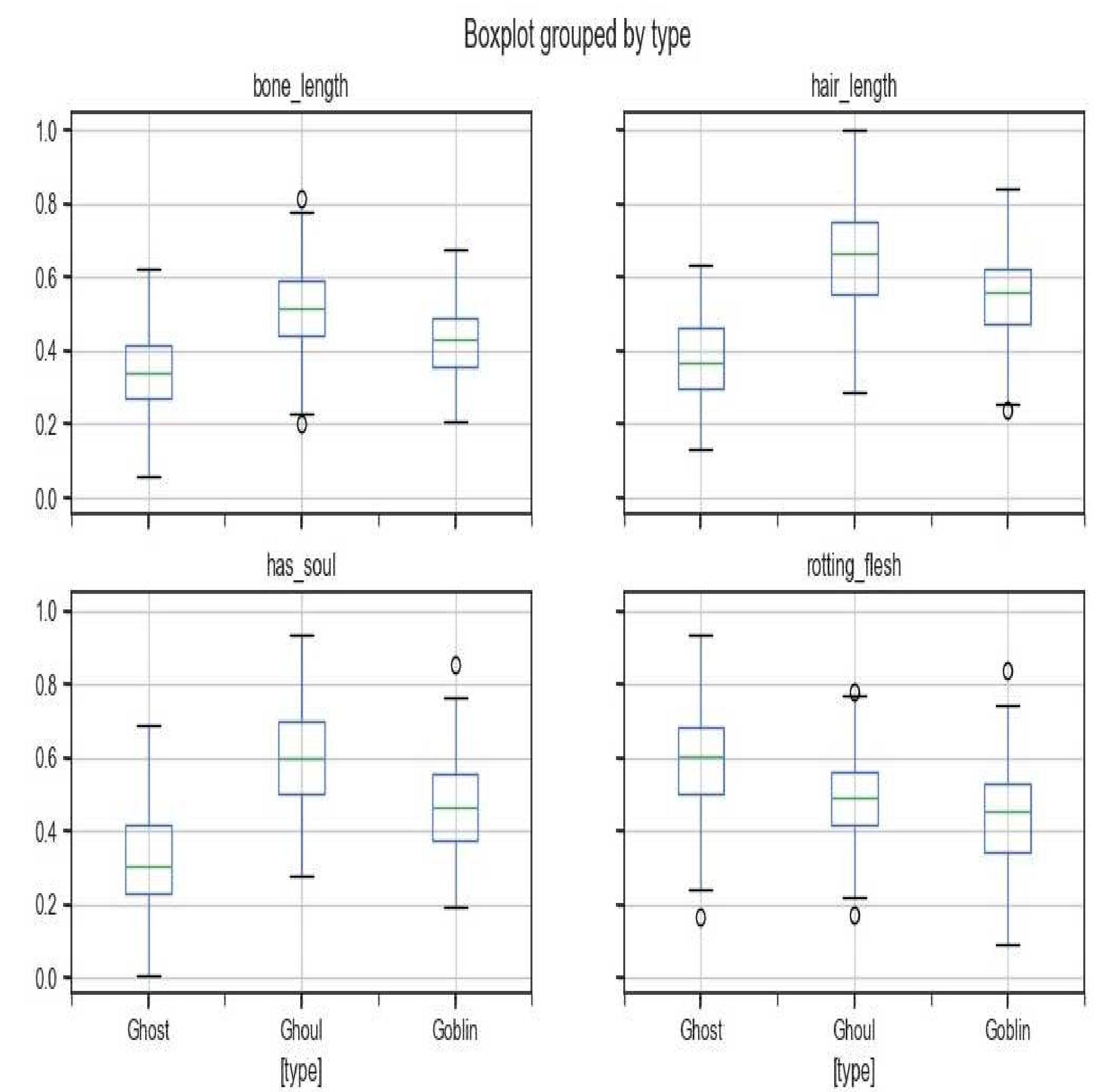


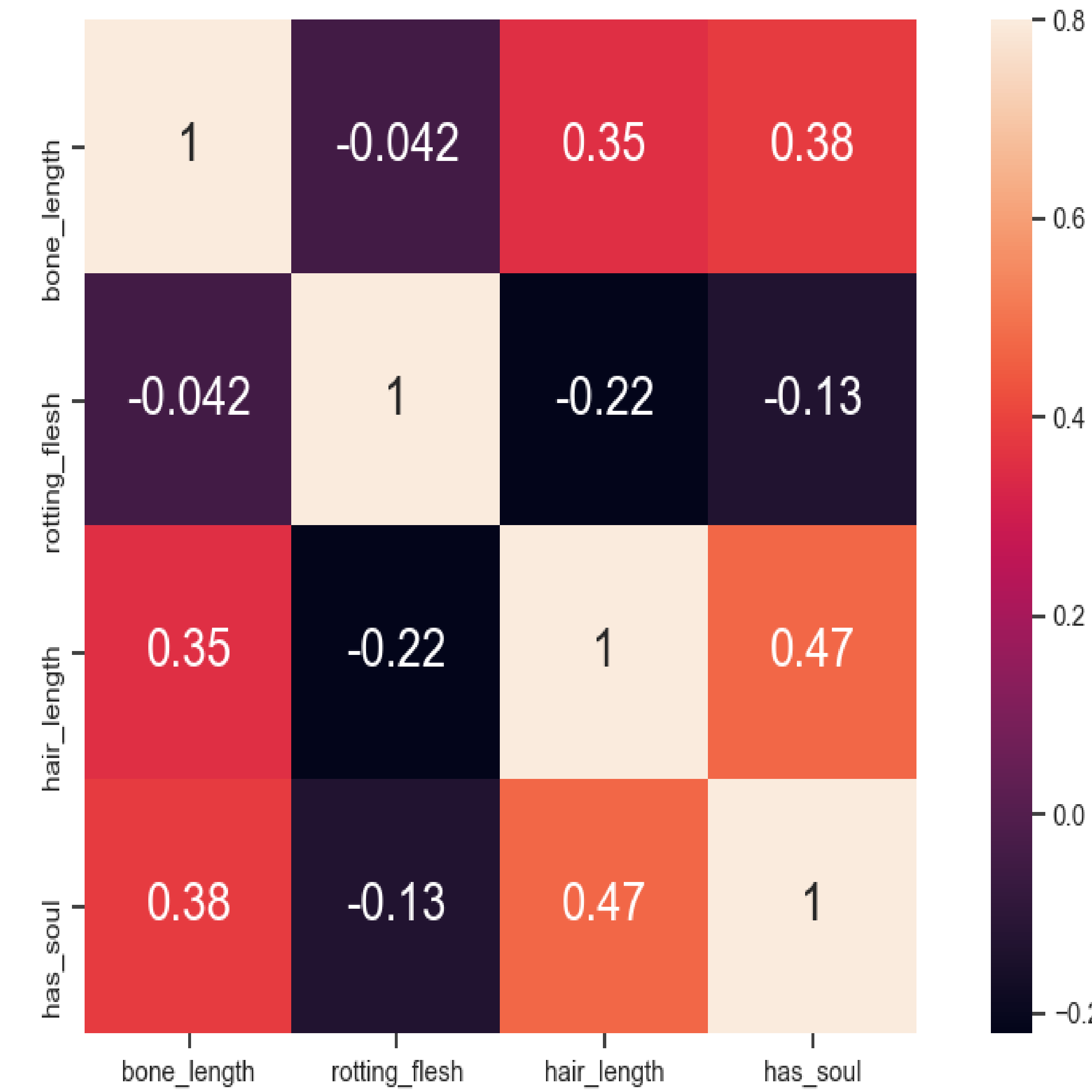
Data Visualization

Problem Statement
Exploratory Data Analysis
Data Visualization
Data Engineering
Methods
Forecast Results
Conclusion
Thanks for attending and welcome for questions

Exp

Based on the above observation on boxplot, we guess that the predictive accuracy of Ghost and Ghoul will be better than Goblin. And the outliers are very small, which can be ignored. As for correllogram, can find that it is no obvious linear relationship between these variables.

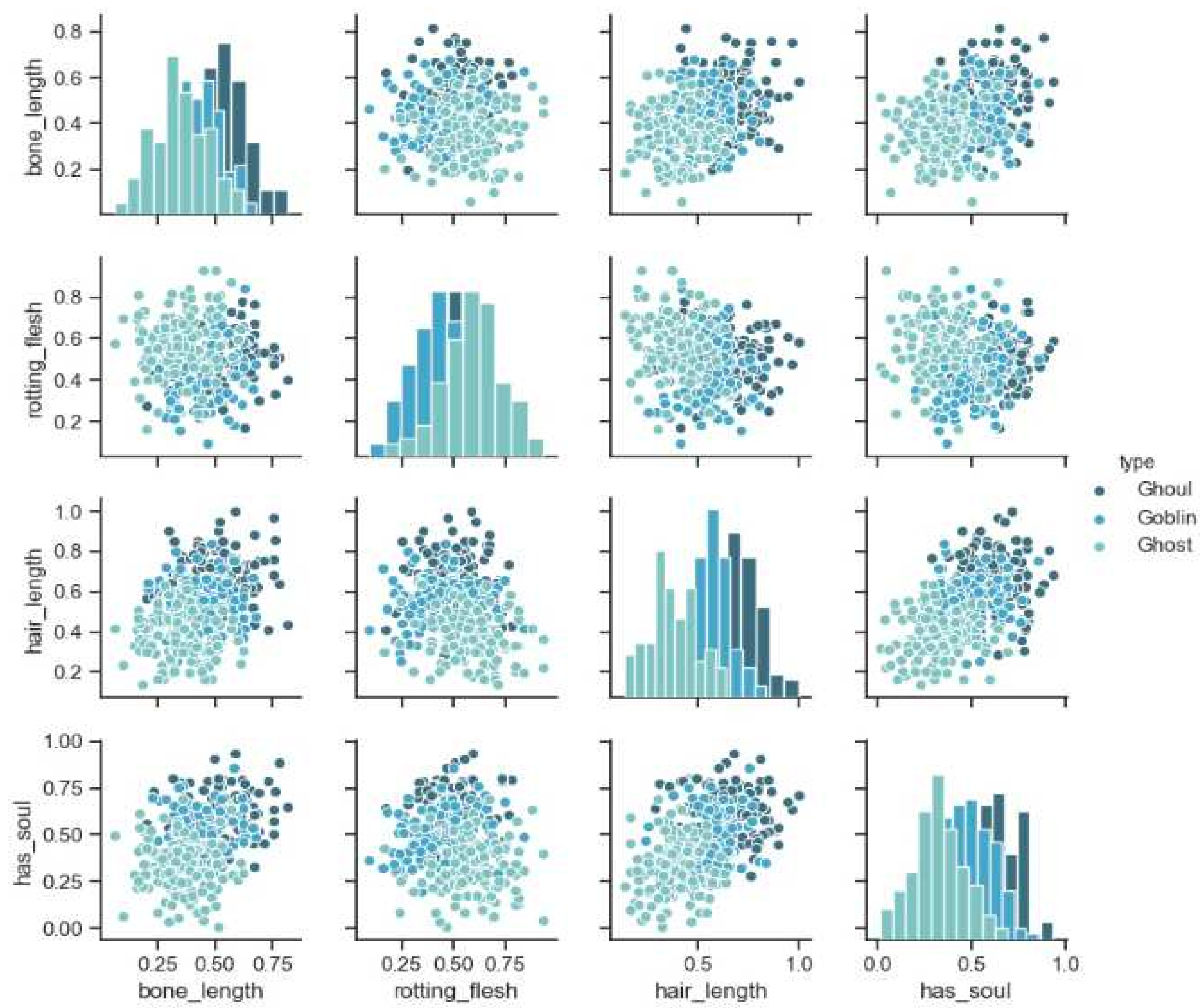




Data Visualization

Exp

This pairplot shows that data is distributed normally. And while most pairs are widely scattered (in relationship to the type), some of them show clusters: hair_length and has_soul, hair_length and bone_length. So it may need to reassemble the data.



Problem Statement
Exploratory Data Analysis
Data Visualization
Data Engineering
Methods
Forecast Results
Conclusion
Thanks for attending and welcome for questions

As it can be seen from the pairplot front the data is distributed normally. But some of them show clusters: hair_length and has_soul, hair_length and bone_length. So create new variables with multiplication of these columns:

- New Features

  hair_soul row[hair_length]*row[has_soul]
  hair_bone row[hair_length]*row[bone_length]
  bone_soul row[bone_length]*row[has_soul]
  hair_soul_bone row[hair_length]*row[has_soul]*row[bone_length]

Features Selection

Problem Statement

Exploratory Data Analysis

Data Visualization

Data Engineering

Methods

Forecast Results

Conclusion

Thanks for attending and welcome for questions

Require: Features $X = \{X_1, X_2, ..., X_n\}$, The number of tree node M, GI_m Gini index of node m, K the number of target, $p_m k$ proportion of target k in node m, $VIM_{jm}^{(Gini)}$ the importance of feature X_j in node m , n the tree number of RF.

Ensure: Variable Importance Measures $VIM_j^{(Gini)}$.

```
Initialize GI_m , VIM_j^(Gini);
for m ← 1...M do
    for k ← 1...K do
        Compute the Gini index of node m GI_m = Σ_{k=1}^{|K|} Σ_{k'≠k} p_mk p_mk' = 1 − Σ_{k=1}^{|K|} p_mk^2
    end for
    Divide node m into node r and node l
    Compute the importance of feature X_j in node m VIM_jm^(Gini) = GI_m − GI_l − GI_r
end for
for i ← 1...N do
    Compute variable importance measures VIM_j^(Gini) = VIM_j^(Gini) + VIM_ij^(Gini)
end for
return VIM_j^(Gini)
```

Line 4: $GI_m = \sum_{k=1}^{|K|} \sum_{k' \neq k} p_{mk} p_{mk'} = 1 - \sum_{k=1}^{|K|} p_{mk}^2$

Line 7: $VIM_{jm}^{(Gini)} = GI_m - GI_l - GI_r$

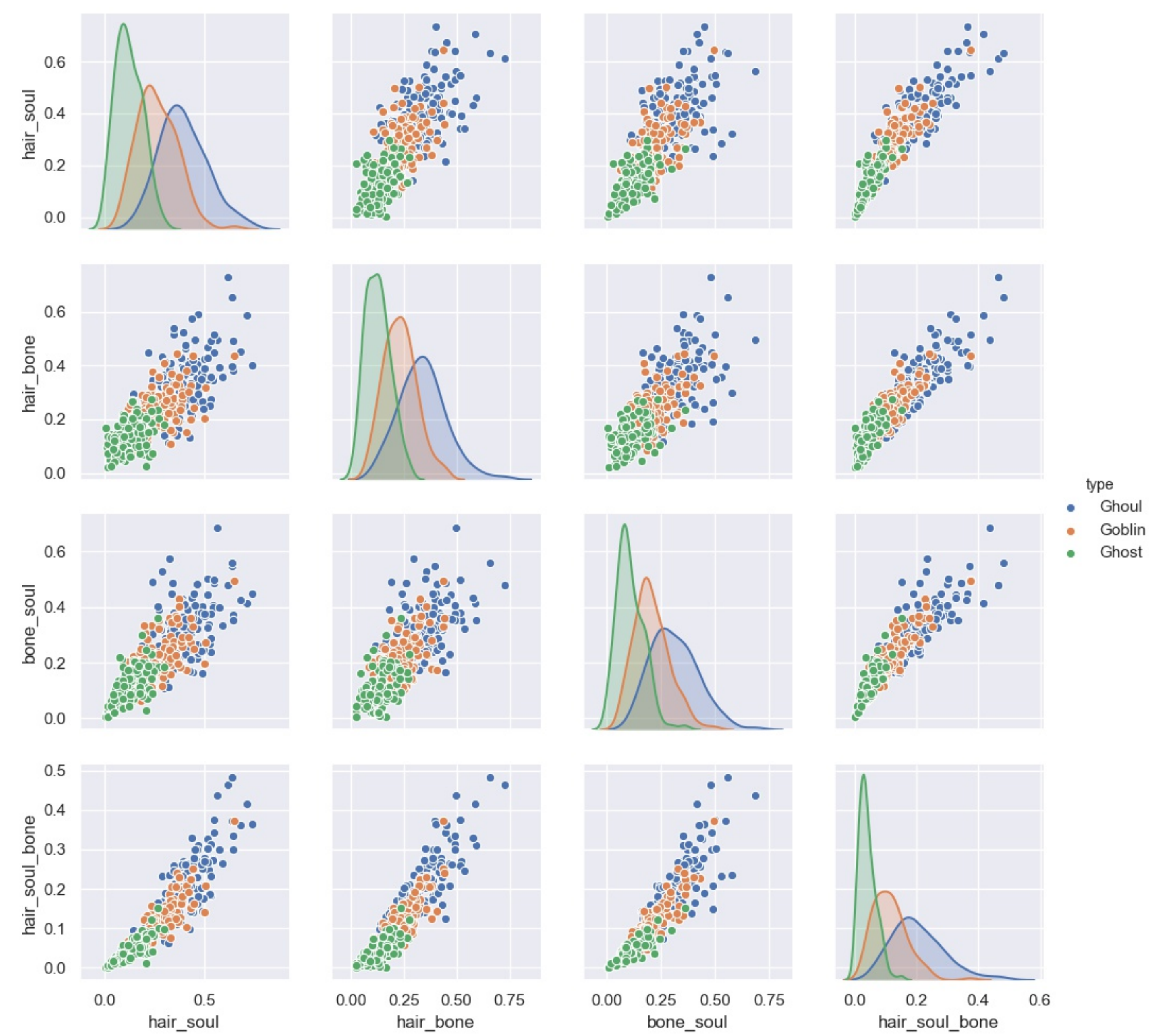
Line 10: $VIM_j^{(Gini)} = VIM_j^{(Gini)} + VIM_{ij}^{(Gini)}$

Algorithm 1: Features Selection

Problem Statement
Exploratory Data Analysis
Data Visualization
Data Engineering
Methods
Forecast Results
Conclusion
Thanks for attending and welcome for questions

Data Engineering

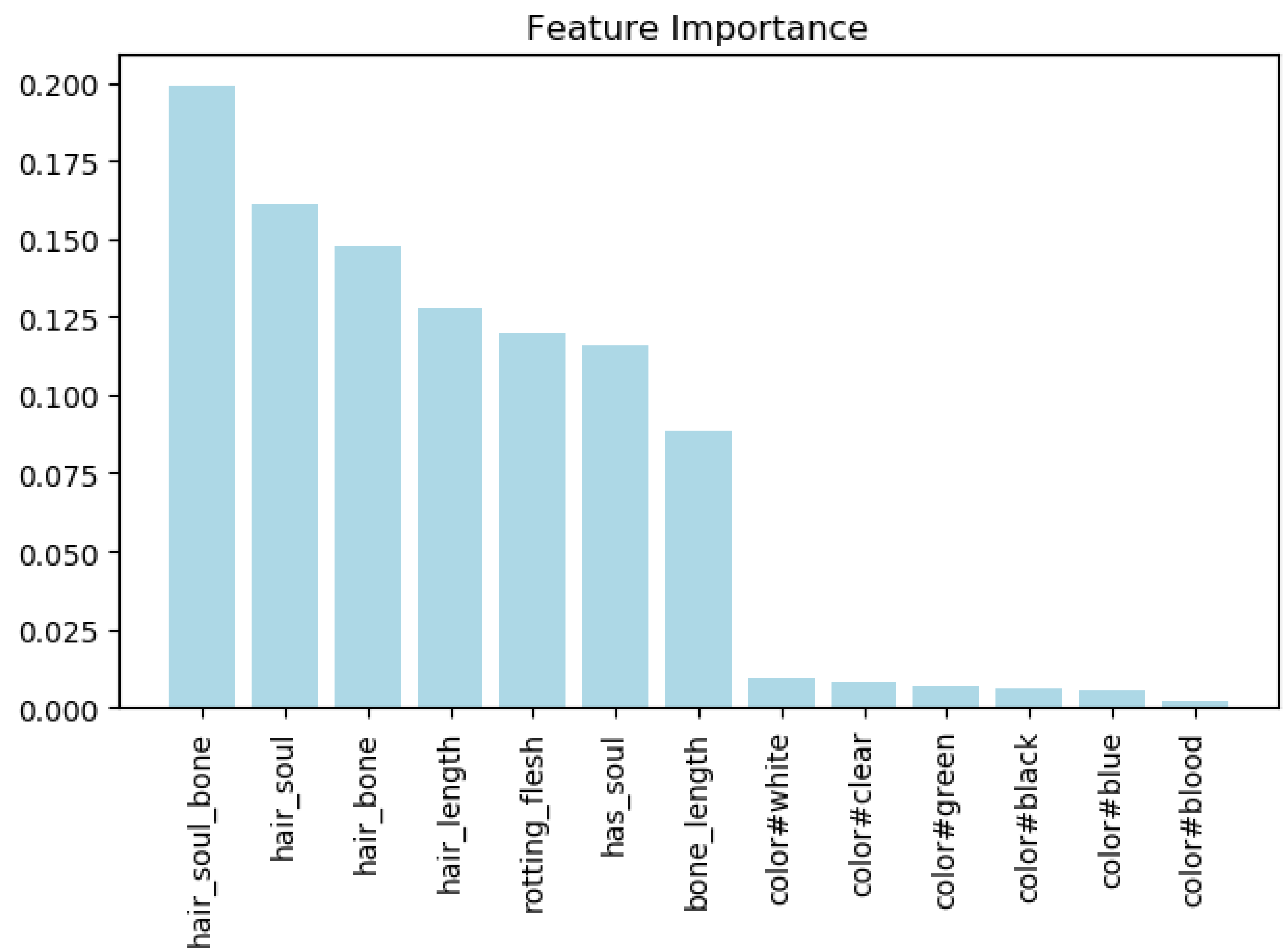
Analyse the new features in a pairplot, it can be seen from the picture that there is a clear linear relationship between the variables.



New Train Data

Problem Statement
Exploratory Data Analysis
Data Visualization
Data Engineering
Methods
Forecast Results
Conclusion
Thanks for attending and welcome for questions

The following figure is a histogram ordered by feature importance. We take the top seven features with higher importance to form a new train data, the rest are discarded.



Problem Statement
Exploratory Data Analysis
Methods
Forecast Results
Conclusion
Thanks for attending and welcome for questions

Methods

Methods

Problem Statement
Exploratory Data Analysis
Methods
Forecast Results
Conclusion
Thanks for attending and welcome for questions

There are many machine learning algorithms, use the machine learning algorithms below as Ensemble Model's base models. Through Grid Search and ten-fold cross-validation to find the optimal parameters. Then use the ensemble model on test data.

- Base Models
  - RandomForeset
  - LogisticRegression
  - SVC
  - KNeighbors
  - XGBoost
  - Netual Network
- Ensemble Model

Problem Statement
Exploratory Data Analysis
Methods
Forecast Results
Evaluation Methods
Forecast Results
Conclusion
Thanks for attending and welcome for questions

Forecast Results

Evaluation Methods

Problem Statement
Exploratory Data Analysis
Methods
Forecast Results
Evaluation Methods
Forecast Results
Conclusion
Thanks for attending and welcome for questions

- F1 Score
- Precision
- Recall

Problem Statement
Exploratory Data Analysis
Methods
Forecast Results
Evaluation Methods
Forecast Results
Conclusion
Thanks for attending and welcome for questions

The tables below are the metrics classification report of ensemble model in original and new train data.

- Metrics Classification Report of Ensemble Model in original and new train data

	data	precision	recall	f1-score	support
Ghost	original	0.80	0.83	0.82	24
	new	0.84	0.88	0.86	24
Ghoul	original	0.88	0.79	0.84	29
	new	0.93	0.97	0.95	29
Goblin	original	0.67	0.73	0.70	22
	new	0.80	0.73	0.76	22
weighted avg	original	0.79	0.79	0.79	75
	new	0.86	0.87	0.86	75

It can be observed that ensemble model performaces better in new features.

Problem Statement
Exploratory Data Analysis
Methods
Forecast Results
Conclusion
Thanks for attending and welcome for questions

Conclusion

Conclusion

Problem Statement
Exploratory Data Analysis
Methods
Forecast Results
Conclusion
Thanks for attending and welcome for questions

Exploratory Data Analysis It is an exploratory analysis of the data to provide the necessary conclusions for data processing and modeling.

Data Preprocessing This step contains dealing with missing data and outliers, changing categorical variable into one-hot code and so on.

Feature Engineering It's the most important thing. Create as more as poosible features, then select the most useful features.

Model Training The models have many parameters, and can use Grid Search to find the optimal paratemers.

Problem Statement
Exploratory Data Analysis
Methods
Forecast Results
Conclusion
Thanks for attending and welcome for questions

Thanks for attending and welcome for questions

Contact Information

Student Shuxia Lin

School of Information Technology

SouthEast University, China

SHUXIALIN@TULIP.ORG.AU