TEXAS
The University of Texas at Austin

Nasa Near-Earth Objects

Using Data to warn us about the Future

Tanvi Dalal, Mark Moreno, Boran Sheu, Parthiv Borgohain
The University of Texas at Austin

Agenda

- About the dataset
- Exploratory Analysis
- Solutions and Insights
- Conclusion and next steps

Problem Statement:

Can we predict through certain variables whether a certified asteroid will be hazardous or not?

Warning!!!

About the Dataset

- 90836 rows
- 10 columns
- 6 variables

	id	name	est_diameter_min	est_diameter_max	relative_velocity	miss_distance	orbiting_body	sentry_object	absolute_magnitude	hazardous
0	2162635	162635 (2000 SS164)	1.198271	2.679415	13569.249224	5.483974e+07	Earth	False	16.73	False
1	2277475	277475 (2005 WK4)	0.265800	0.594347	73588.726663	6.143813e+07	Earth	False	20.00	True
2	2512244	512244 (2015 YE18)	0.722030	1.614507	114258.692129	4.979872e+07	Earth	False	17.83	False
3	3596030	(2012 BV13)	0.096506	0.215794	24764.303138	2.543497e+07	Earth	False	22.20	False
4	3667127	(2014 GE35)	0.255009	0.570217	42737.733765	4.627557e+07	Earth	False	20.09	True

TEXAS
The University of Texas at Austin

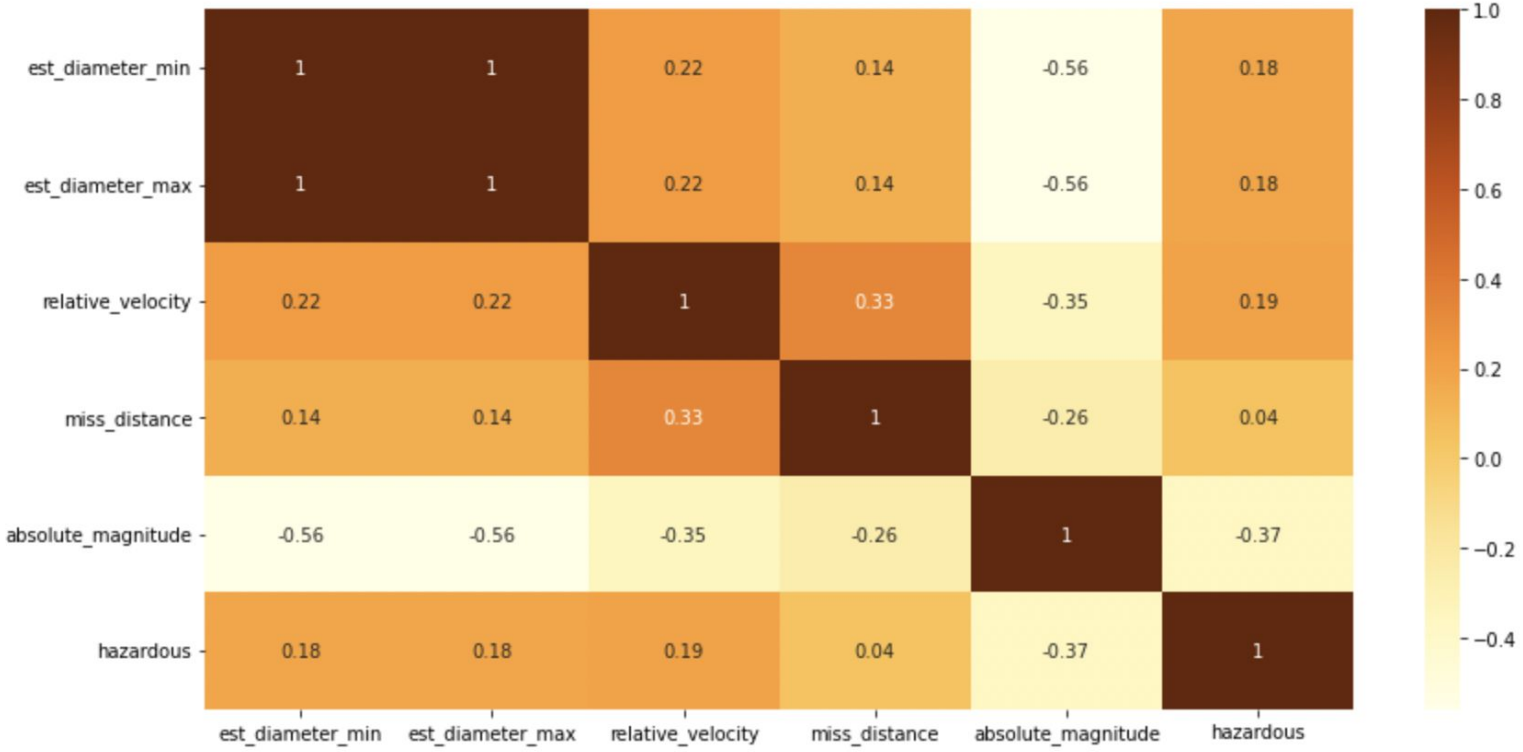
Exploratory Data Analysis

Variables Analysis & Cleaning

- Estimated diameter min and estimated diameter max are correlated so one of them will be removed
- Orbiting body and sentry object will be removed
- Columns like name, id will also be removed

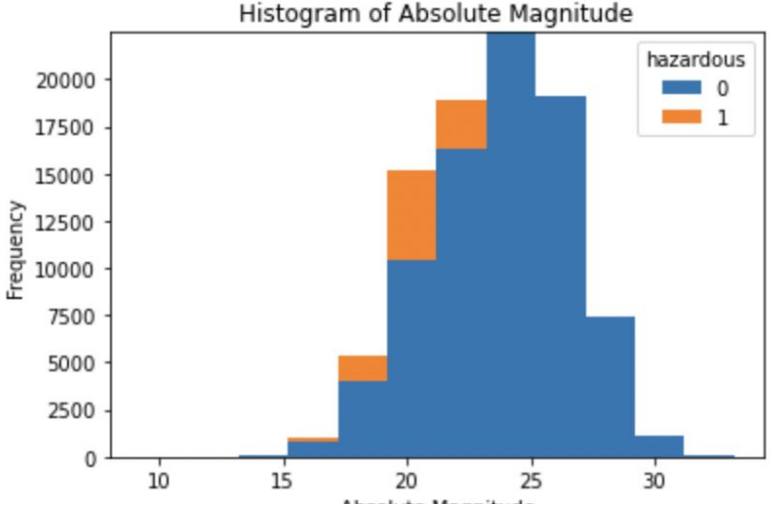
	id	name	est_diameter_min	est_diameter_max	relative_velocity	miss_distance	orbiting_body	sentry_object	absolute_magnitude	hazardous
0	2162635	162635 (2000 SS164)	1.198271	2.679415	13569.249224	5.483974e+07	Earth	False	16.73	False
1	2277475	277475 (2005 WK4)	0.265800	0.594347	73588.726663	6.143813e+07	Earth	False	20.00	True
2	2512244	512244 (2015 YE18)	0.722030	1.614507	114258.692129	4.979872e+07	Earth	False	17.83	False
3	3596030	(2012 BV13)	0.096506	0.215794	24764.303138	2.543497e+07	Earth	False	22.20	False
4	3667127	(2014 GE35)	0.255009	0.570217	42737.733765	4.627557e+07	Earth	False	20.09	True

Correlation between the Variables of interest



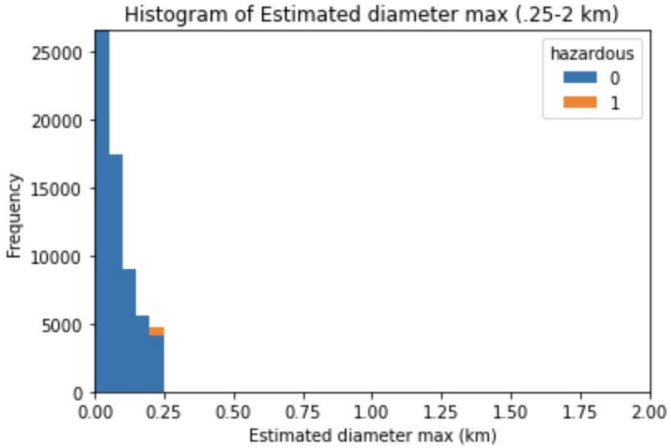
Absolute Magnitude

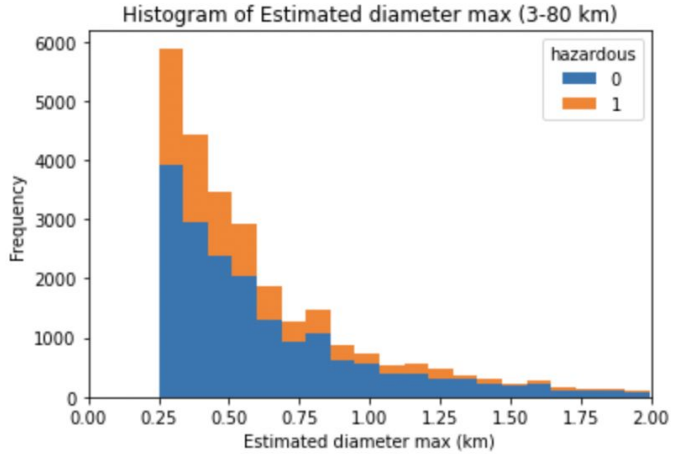
A measure of luminescence used for measuring the approximate diameter of an asteroid.



Estimated Diameter

- About 75% of all hazardous are from the range of .25 to 1 kilometer. While only accounting for ⅛ of the total number of asteroids
- Asteroids smaller than .25 kilometers were only hazardous < 1% of the time while those larger were hazardous 40% of the time.



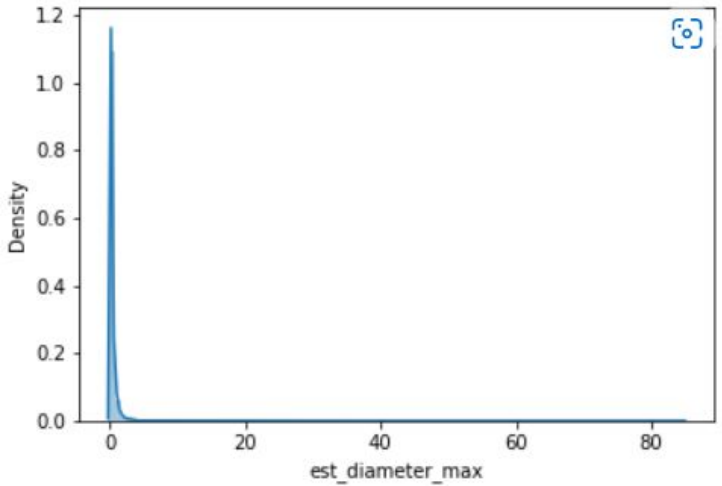


TEXAS
The University of Texas at Austin

Solutions and Insights

Naive Bayes

Qcut: Estimated Diameters



	Positive class	Negative class	Positive/Negative Ratio	Importance
Q("absolute_magnitude_binned_(22.8, 24.5]")	0.000076	0.044950	0.001688	6.384180
Q("absolute_magnitude_binned_(26.1, 33.2]")	0.000076	0.044049	0.001723	6.363918
Q("absolute_magnitude_binned_(24.5, 26.1]")	0.000076	0.043106	0.001760	6.342296
Q("relative_velocity_binned_(203.345, 25865.613]")	0.009561	0.043050	0.222085	1.504694
Q("est_diameter_max_binned_1")	0.008195	0.002234	3.668732	1.299846
Q("absolute_magnitude_binned_(9.229000000000001, 20.82]")	0.115790	0.031865	3.633802	1.290279
Q("est_diameter_min_binned_1")	0.000986	0.000414	2.381213	0.867610
Q("absolute_magnitude_binned_(20.82, 22.8]")	0.084073	0.036040	2.332788	0.847064
Q("relative_velocity_binned_(67870.599, 236990.128]")	0.075271	0.036543	2.059779	0.722599
Q("relative_velocity_binned_(25865.613, 37650.841]")	0.025799	0.041222	0.625847	0.468650
Q("miss_distance_binned_(6745.532, 12868712.73]")	0.028454	0.041068	0.692866	0.366919
Q("relative_velocity_binned_(50953.539, 67870.599]")	0.051597	0.039484	1.306797	0.267579
Q("miss_distance_binned_(60281596.401, 74798651.452]")	0.045299	0.039573	1.144701	0.135143
Q("miss_distance_binned_(45371571.727, 60281596.401]")	0.044237	0.039606	1.116940	0.110593
Q("miss_distance_binned_(29989093.462, 45371571.727]")	0.042112	0.039492	1.066358	0.064249
Q("relative_velocity_binned_(37650.841, 50953.539]")	0.037863	0.039711	0.953467	0.047651
Q("est_diameter_max_binned_0")	0.191669	0.197752	0.969239	0.031244
Q("miss_distance_binned_(12868712.73, 29989093.462]")	0.039988	0.040272	0.992954	0.007071
Q("est_diameter_min_binned_0")	0.198877	0.199571	0.996522	0.003484

High amount of data are relatively small, but the range is large!
Results in importance actually fits our estimation.

Naive Bayes

Accuracy:

1. Prior Probability for No Hazard: **90.15%**

2. Naive Bayes:
   a. Training data: **89.77%**
   b. Testing data: **89.46%**

F1 Score:

1. Precision score: **27.9%**
2. Recall score: **3.6%**

3. F1 Score
   a. Training data: **6.0%**
   b. Testing data: **6.4%**

Likelihoods→Importance:

	Positive class	Negative class	Positive/Negative Ratio	Importance
Q("absolute_magnitude_binned_(22.8, 24.5]")	0.000074	0.044586	0.001665	6.398180
Q("absolute_magnitude_binned_(26.1, 33.2]")	0.000074	0.044138	0.001681	6.388086
Q("absolute_magnitude_binned_(24.5, 26.1]")	0.000074	0.043373	0.001711	6.370594
Q("relative_velocity_binned_(203.345, 25865.613]")	0.010094	0.042998	0.234743	1.449266
Q("absolute_magnitude_binned_(9.229000000000001, 20.82]")	0.116595	0.032039	3.639154	1.291751

Logistic Regression

Accuracy:

-Baseline: **0.9024298183533852**

-Prediction: **0.9032696047851455**

```
relative_velocity    -1.616650e-05
miss_distance        -2.987473e-08
absolute_magnitude   -2.242601e-08
est_diameter_min      9.423904e-11
est_diameter_max      2.107249e-10
dtype: float64
```

Variable Importance:

- Most significant:
  - **est_diameter_min**
  - **est_diameter_max**
- Least significant:
  - **relative_velocity**

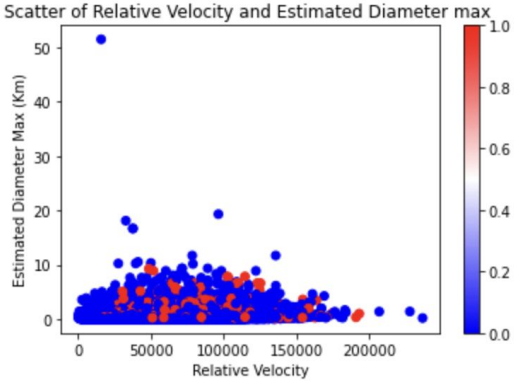
K- Nearest Neighbors

Baseline:

0.9024298183533852

Prediction:

0.9031228211808741




```
array([[24606,     9],
       [ 2631,     5]])
```

F-Score = .0006

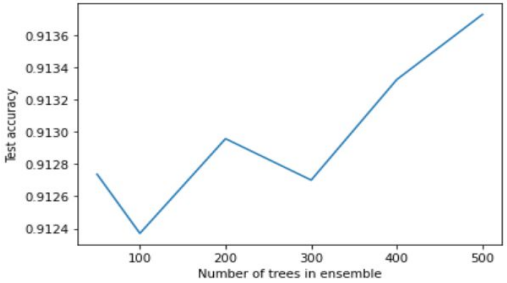
Trees and Ensemble Methods

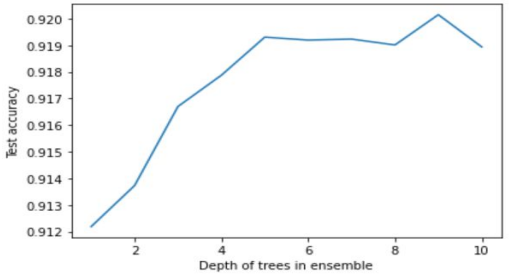
- First we a tried to fit decision trees and other ensemble classifiers like Bagging, Random Forest and Gradient Boosting without playing too much with the parameters.
- The metrics obtained via these models are shown on the right.
- Clearly, the big decision Tree was **overfitting with a training accuracy of 100%** and a t**est accuracy of 89.4%** (even lesser than the baseline accuracy).
- Even **Bagging and Random Forest were overfitting** with much higher accuracy values for the training datasets compared to the test datasets.
- **Boosting** seemed to be performing the **best on the Test Set.**

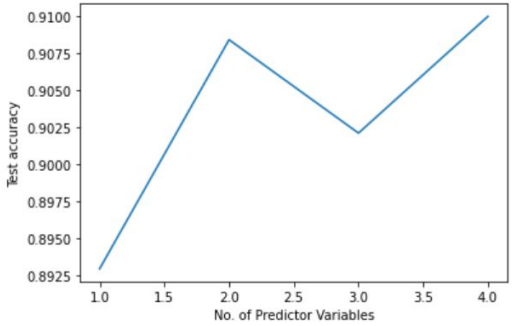
	Training Accuracy	Test Accuracy	Train F1 Score	Test F1 Score
Decision Tree	1.000000	0.893765	1.000000	0.461996
Bagging	0.991570	0.912150	0.955003	0.430813
Random Forest	0.962507	0.908444	0.771033	0.375469
Gradient Boosting	0.922120	0.915232	0.379760	0.325350

Trees and Ensemble Methods

- Next, we will try to tune the parameters for Gradient Boosting.
- On varying the no. of estimators across 100,200,300,400 and 500, we obtained the plot on the right. The differences are minuscule. We will take no. of trees = 500
- On varying the maxdepth from 1 to 11, we obtained the second plot on the right. The test accuracy seems to be maxing out at maxdepth=9.
- Finally, we will try to find the best value for the no. of predictors to take for random forest. On varying the no. of predictors from 1 to 4, we obtain the third plot on the right. Hence we will take mtry=4 for our random forest model.





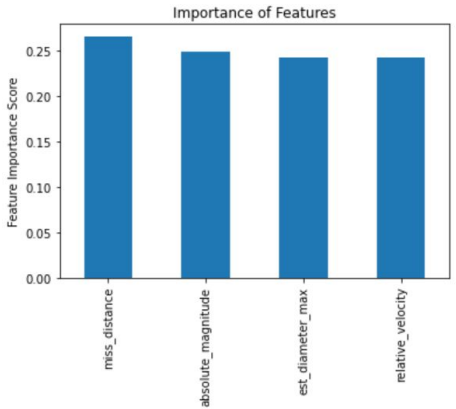


Trees and Ensemble Methods

- On running our models with the parameters decided in the previous slide, we obtain the metrics given on the table in the right.
- Clearly, Gradient Boosting is performing the best with **Test Accuracy** of **92.02%** (compared to **90.2% baseline accuracy**). The **F1 Score** obtained is **0.48**
- From the confusion matrix on the right, **Precision = 0.65** and **Recall = 0.38** which is definitely an improvement over the baseline model.
- On plotting the feature importance for the Gradient Boosted Decision Tree Model, we get the third figure on the right.
- Clearly, nearly all the variables have similar importances with no single dominant variable. Among the 4 variable, **miss_distance** seems to be the **most important feature**.

	Training Accuracy	Test Accuracy	Train F1 Score	Test F1 Score
Decision Tree	1.000000	0.892701	1.000000	0.459519
Bagging	0.991570	0.912150	0.955003	0.430813
Random Forest	0.977888	0.910022	0.874911	0.413678
Gradient Boosting	0.986978	0.920150	0.928448	0.480668

```
array([[24068,   531],
       [ 1645,  1007]], dtype=int64)
```



TEXAS
The University of Texas at Austin

Conclusion and Next Steps

Conclusion

1. The models being used by us do not seem to be extremely good at predicting whether an NEO will be hazardous or not. We have achieved only modest gains over the baseline accuracy
2. Only **4 variables** are being used to predict **whether an NEO is hazardous or not** (**miss_distance, absolute magnitude, est_diameter_max and relative_velocity**).
3. None of the 4 selected variables seem to be heavily impacting the target variable. All the **4 variables seem to have similar values for feature importance** in the Gradient Boosted Tree Model. Other classification models too did not give a wide variation in relative importance of features.

Conclusion

1. Given the aforementioned point, it seems reasonable to expect the findings of Point 1 as the four variables do not seem to be great predictors. Our best model obtained a **test accuracy of 92.015%** and an **F1 score of around 0.48**
2. **miss_distance** seems to be the **most important feature** as per our best model (**Gradient Boosted Decision Tree**). However, as per **Logistic Regression**, this is **est_diameter_max** and as per **Naive Bayes** it is **absolute_magnitude**. This variance probably probably suggests that no strong relationship exists and all of the variables are nearly similar in importance.

Next Steps

- Find a larger dataset with more even dependent variable distribution
- Use more variables
- Limitation: many asteroids were listed multiple times

TEXAS
The University of Texas at Austin

Questions?