**Admission Analyzer**

Adil Mohammad

Ashutosh Joshi

Rahul Kaushik Mallela

Rohit Kumar Koutike

UMBC DATA 604

Dr. Sarika Rajeev

**ABSTRACT**

Trying to figure out what universities to apply to can be quite duanting. The internet has a plethora of information available but it can be easy to get overwhelmed by the large amounts of information. This can lead to paralysis by analysis, leading to a further delay in the application process. Therefore, we thought it would be a great idea to focus on making an application that can give an output of universities and their admission chances percentage. All the user would have to input would be GRE, TOEFL, LOR, SOP, and research paper Experience and University tier ranks. By implementing this solution people can reduce search time and obtain useful information from raw data. This paper will be helpful to anyone who wants to study abroad but is unsure of which university to apply to or the university admission they will receive. The paper goes on to give a brief introduction, displays visualizations of our dataset, showcases our prototype, and discusses the steps we would take to deploy this model into a live application.

**INTRODUCTION**

A wide range of factors are taken into consideration when ranking an applicant's requirements for a particular program during the process of graduate school admission. Test results, grades, research experience, letters of reference, a statement of purpose, and other pertinent accomplishments are typically considered in combination with academic and non-academic criteria while making admission selections. In our research, we provide students who are thinking about applying to universities with a complete dataset that contains data on the factors that affect graduate school admissions in the US.

The dataset can give students information about the admissions process, help them in selecting where to apply and how to present themselves in their applications, and ultimately increase their chances of getting accepted into the program of their choice based on their university ranking and test results. Here, we employed machine learning methods like the linear regression model and the Spark SQL methodology to predict and understand the graduate school admissions process. Students can find patterns and correlations that can assist students forecast their chances of being admitted to a specific program at a specific university by studying big datasets and collecting data on every aspect taken into account during the admissions process.

The methodology allows students to estimate their chances of getting into a US graduate program and make informed application decisions. Colleges can use the dataset to examine the requirements for admission to their graduate programs and enhance the accuracy of their admissions decisions. The Budget Prediction outlines the financial requirements for students to register or begin the application process using our Web application.

**LITERATURE SURVEY**

* [**https://www.topuniversities.com/where-to-study/north-america/united-states/ranked-top-100-us-universities**](https://www.topuniversities.com/where-to-study/north-america/united-states/ranked-top-100-us-universities)
* [**https://aws.amazon.com/ec2/instance-types/**](https://aws.amazon.com/ec2/instance-types/)
* [**https://aws.amazon.com/aws-cost-management/**](https://aws.amazon.com/aws-cost-management/)
* [**https://spark.apache.org/docs/1.0.2/sql-programming-guide.html**](https://spark.apache.org/docs/1.0.2/sql-programming-guide.html)

**OBJECTIVE**

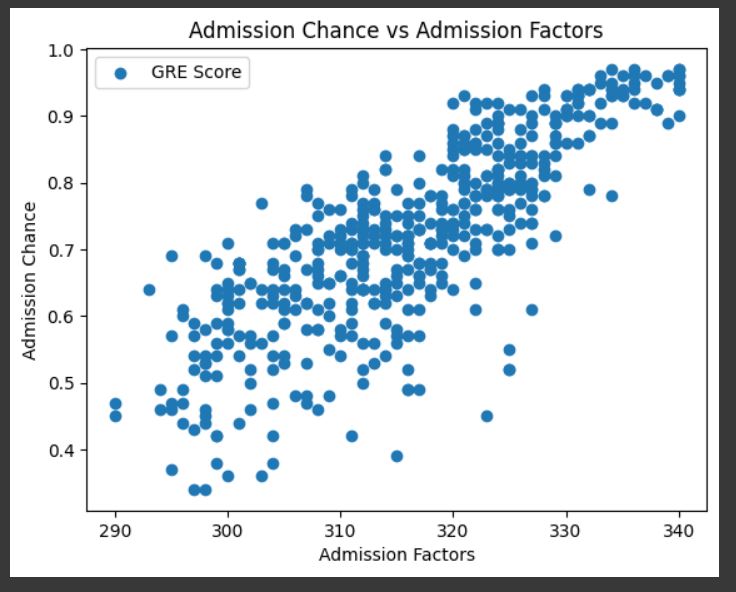
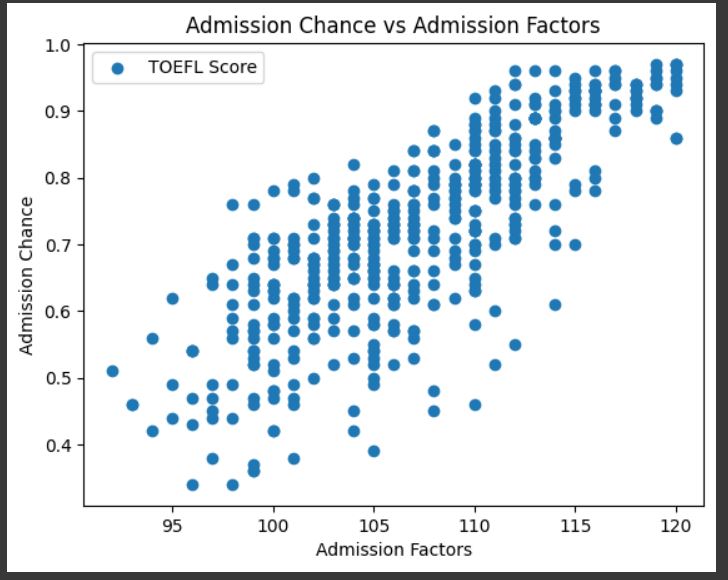
Our objective is to create an application where international students can input their GRE score, TOEFL, SOP, LOR, CGPA, and research experience in order to find out their chances of getting into a graduate university. The output will show a list of universities ranked with the highest number of percentages. For this project we focused on using a prototype model that was trained using a dataset from a csv file. This project gives us insight on how we would approach this project on a larger scale.

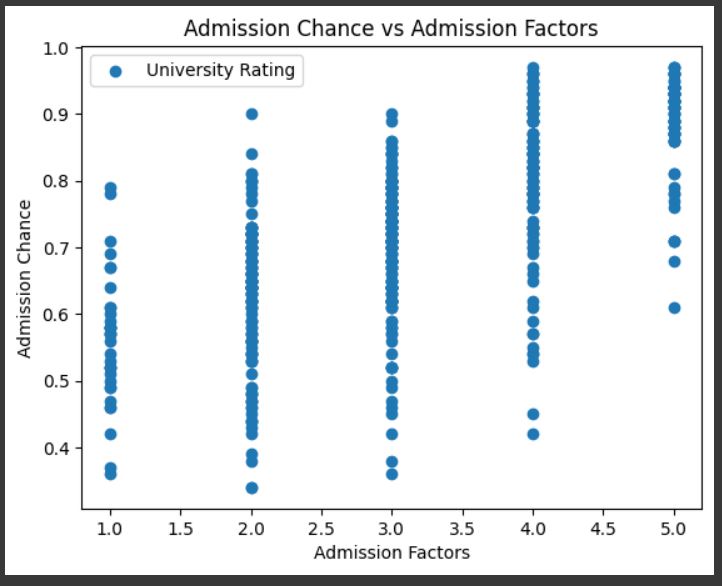
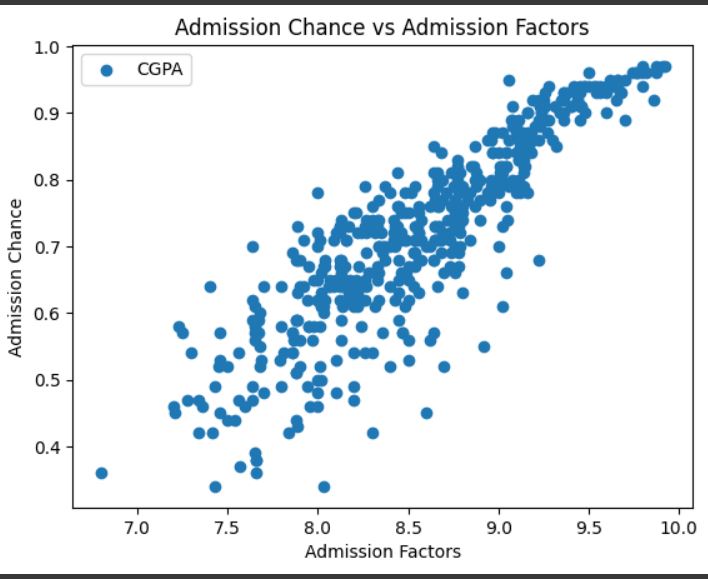
**DATASET**

The dataset we have described contains valuable information on various factors that are considered important for admissions to graduate schools in the US. The dataset includes several columns such as GRE score, TOEFL score, university rating (tiers), SOP score, LOR score, CGPA, and Research experience score, which are all important indicators of a student's academic and research capabilities. By analyzing this dataset, students can gain insights into the admissions process for graduate schools and can use this information to predict their chances of being accepted into a program.

**VISUALIZATIONS**

In order to gain more insight into our dataset, we made these visualizations to see the correlation between admission chances and admission factors.





**Correlations**



We found that the GRE scores and GPA had the highest correlations. Through this analysis we had analyzed that these were the strongest indicators. In our modeling we would have to leverage the use of all these features in order to have a suitable model that we could put into production.

**Our Model**

For this dataset we implemented the use of multivariate linear regression and we were pleased with an accuracy of 96%.

------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------

**Next Steps needed to implement our project on a larger scale to put into the marketplace**

**Data Acquisition**

Obviously, if we wanted to put this model into production, it would not be sufficient to only run a model off of one csv file. In order to make our model applicable, we would need to collect our data through a variety of sources. One way we could do this is through surveys which would be sent out to graduate students that ask about their academic performances and university choices. Our most optimal source of data collection would be by scraping the data from university data and publicly available data.

**Tools Needed to Reach our Objective**

**Python, Spark, SQL, and AWS**

****

**Steps Needed to Reach our Objective**

**1)Data Acquisition:**

* As mentioned above, we would use those techniques in order to gather data. In order to do that we could leverage the use of PySpark and/or Spark SQL by using Spark’s distributed computing capabilities to scrape large amounts of data.

**2)Data Preprocessing/Feature Engineering:**

* Clean and preprocess the collected data to ensure consistency and remove any irrelevant or incomplete information by using Spark SQL.
* Merge the data from different sources based on common identifiers, such as university names or program names by using Spark SQL.
* Transform any categorical variables into numerical or binary representations that can be used by the model Spark SQL.
* We can use PySpark’s Dataframe API in order to feature engineer the certain parameters that will help our model perform better.

**3) Model Training:**

* We would try to use multiple models and compare their metrics in order to see which model performs best. The models we have in mind are multivariate linear regression, random forest, and K-nearest neighbors. We would train and test these models using PySpark and even Spark SQL depending how big the dataset is.
* Then, we would train our model using the training data and tune its hyperparameters to optimize its performance. Cross validation and grid search would help us find the optimal hyperparameters.

**4)Model Evaluation:**

* Evaluate our model's performance using appropriate metrics, such as accuracy, mean squared error, or ranking metrics like NDCG (Normalized Discounted Cumulative Gain) or MAP (Mean Average Precision).

**5)Deployment:**

* Once we are satisfied with the performance of our best model, we would be ready to use AWS to deploy our model and also host our website. AWS EC2 instances would be used for model deployment and AWS Lambda would be helpful in having a serverless deployment of the model.

**Budget Analysis**

The budget to industrialize this project requires following requisites:

Amazon Web Services:

t2.large 2vCPU, 8 GB RAM, EBS Storage: 36 CPU Cred/ Hour

0.26 $ per hour. Running for a month

=> 0.26 \* 24 \* 30 = **187.2 $ per month**

Maintenance:

**100-120 $ per month**

Web Hosting:

**30-40 $ per year**

Total: 187 $ + 110 $ + 5 $ => **302 $ per month**

With a budget of under $500 per month we can deploy this project in the market

**CONCLUSION**

The goal of technology is to make people’s lives simpler. Machine learning has revolutionized every industry and will continue to do so. We have proposed a machine learning solution for an audience that has a huge sample size. In order to accurately predict people’s chances of getting admission into a university, we must process huge loads of data. This class has taught us how it is imperative that we use assistance from platforms such as Spark and AWS in order to optimize our solution. In the future, we would like to add provisions to this application to add more features. An example of an enticing feature could be a feature that takes into account a student’s budget for college tuition. Eventually, we would like to coordinate with the universities themselves in order to improve and simplify the admissions process for students. All in all, we believe that this project gave us great insight into machine learning, data management, and big data platforms.