Fortnightly Task 1

Simon Karumbi
S3455453

Task 1: Job Role

Find a job advertisement for a data science position that covers the type of data science role you see yourself doing

1.1a

Include a copy of the job description.

https://boards.greenhouse.io/doordash/jobs/3219350?gh_jid=3219350

Data Scientist, International Analytics at DoorDash
MELBOURNE, VICTORIA, AUSTRALIA
About the Role
The Analytics team is looking for experienced Data Scientists to guide measurement, strategy, and tactical decision-making as we expand our logistics platform into new countries. We are looking to hire in several countries across a variety of teams and levels to accelerate growth in our current international markets (Canada and Australia) and lead the strategy and execution of new market launches (more coming!) Data Scientists at DoorDash work to uncover insights and turn them into relevant recommendations, driving decisions for the entire organization. Analytics is integral to all operational areas at DoorDash.

As a Data Scientist at DoorDash, you'll use your quantitative background to mentor other scientists and dive into large datasets to guide decision-making. We tackle a multitude of exciting challenges including customer acquisition, balancing supply and demand, fraud and support, marketing, marketplace efficiency, and more. If you enjoy finding patterns amidst chaos, are excited to build a market from 0 to 1, and have experience using analytics to affect revenue, growth, operations or beyond, we're looking for someone like you!

You're excited about this opportunity because you will…
Use quantitative analysis and the presentation of data to see beyond the numbers and understand what drives our business

Build full-cycle analytics experiments, reports, and dashboards using SQL, R, Python, or other scripting and statistical tools
Produce recommendations and use statistical techniques and hypothesis testing to validate your findings
Provide insights to help business and product leaders understand marketplace dynamics, user behaviors, and long-term trends
Identify and measure levers to help move essential metrics and make recommendations
Work backwards from understanding and sizing problems to ideating solutions
Report against our goals by identifying essential metrics and building executive-facing dashboards to track progress
Be excited to travel (when it's safe!) to meet with business partners and the team in each market
We're excited about you because you have…
A degree in Math, Physics, Statistics, Economics, Computer Science, or similar domain
5+ years of experience in data analytics, consulting, or related quantitative role
Experience working with funnel optimization, user segmentation, cohort analyses, time series analyses, regression models, etc
Expertise of SQL queries, ETL, A/B Testing, and statistical analysis (e.g. hypothesis testing, experimentation, regressions) with statistical packages, such as Matlab, R, SAS or Python
Proficiency in one or more analytics & visualization tools (e.g. Chartio, Looker, Tableau)
The insight to take ambiguous problems and solve them in a structured, hypothesis-driven, data-supported way
The determination to initiate and lead projects to completion in a scrappy environment
Prior experience working abroad or in international expansion preferred but not required
Fluent English required, proficiency in additional languages a plus
Why You'll Love Working at DoorDash
We are leaders - Leadership is not limited to our management team. It's something everyone at DoorDash embraces and embodies.
We are doers - We believe the only way to predict the future is to build it. Creating solutions that will lead our company and our industry is what we do -- on every project, every day.
We are learners - Everyone here is continually learning on the job, no matter if we've been in a role for one year or one minute.

We are customer-obsessed - Our mission is to grow and empower local economies. We are committed to our customers, merchants, and dashers and believe in connecting people with possibility.
We are all DoorDash - The magic of DoorDash is our people, together making our inspiring goals attainable and driving us to greater heights.
We offer great compensation packages and comprehensive health benefits.

About DoorDash
DoorDash is a technology company that connects customers with their favorite local and national businesses in all 50 US states, Canada, and Australia. Founded in 2013, DoorDash empowers merchants to grow their businesses by offering on-demand delivery, data-driven insights, and better in-store efficiency, providing delightful experiences from door to door. By building the last-mile delivery infrastructure for local cities, DoorDash is bringing communities closer, one doorstep at a time. Read more on the DoorDash Engineering blog or at www.doordash.com.

Our Commitment to Diversity and Inclusion
We're committed to growing and empowering a more inclusive community within our company, industry, and cities. That's why we hire and cultivate diverse teams of the best and brightest from all backgrounds, experiences, and perspectives. We believe that true innovation happens when everyone has room at the table and the tools, resources, and opportunity to excel.

1.1b

This role pertains to food delivery and relates to customer service and satisfaction.

1.1c

The role seems to be more directed at gathering insights from data, due to the extensive usage of the term analytics. In saying this, as the role is looking at future growth in unknown markets, making predictions would be a crucial part of this role. This is supported by the role description which stipulates that insights will be gained through analysis to 'understand what drives our businesses' with predictions through statistical inference also playing a large part of this role, where 'statistical techniques and hypothesis testing' will be made to validate findings.

Task 1.2: Data Set

1.2a

https://www.kaggle.com/c/santander-customer-satisfaction/data

This data set is from the Santander Bank in Europe, and contains several anonymised numerical variables and a target column specifying satisfied and unsatisfied customers.

1.2b

This data set was chosen as it addresses a problem that would be comparable between both food delivery and financial institutions, customer satisfaction. After having looked for specific data sets on food delivery, there are not many good quality data sets available online. Although in a food delivery scenario there would be more categorical fields that might be used, the Santander data set might've been preprocessed with categorical variables being numerically encoded.

Task 1.3: Experiment

1.3a

A Support Vector Machine Classifier and a Gaussian Naive Bayes Classifier were used, with differing success. Without performing extensive exploratory data analysis, insights are hard to come by, when purely using Machine Learning Algorithms. The Support Vector Machine made predictions of only 0, indicating that the target class of 0 is overwhelmingly proportioned compared to the other data, whereas the Gaussian Naive Bayes model made some predictions for some of the test cases.

1.3b

The Support Vector Machine achieved an impressive 96% accuracy score, indicating that it might be a suitable Machine Learning Model to tune for better results. In comparison, the Gaussian Naive Bayes Classifier only achieved an accuracy score of 9.7%, indicating that the data may not be distributed normally, and there may be several outliers in the data that are affecting the success of the model.

Accuracy may not be a suitable measure for this classification problem, as it is flawed when used on target classes with unequal proportions, where the Support Vector Machine seemed unusually high. Other measures such as an F1 score could be explored, where it should be considered whether recall or precision are most important measures

for business practice. This means considering whether accurately predicting satisfied or unsatisfied customers is of more importance.

1.3c

Fairness Unaware Analysis is unable to be performed on this dataset, as the feature names are unknown due to their anonymisation.

Part 2: Reflection

2a

The most challenging aspect of this task was the Fairness Aware Analysis of the data analysis. I have not had any experience in using these tools, nor do I yet understand their functionality, however, I am looking forward to learning more about it as I continue my studies. From my research, it seemed that the dataset that I had selected was unsuitable for this analysis as it was impossible to identify where bias might have found its way into the model with anonymised features. Finding a suitable data set was a bit of a challenge, although there are plenty of resources available, where the ‘perfect’ data set for the use case is more elusive than I anticipated and took quite a bit of time. I feel that the data set I did end up selecting was somewhat suitable in terms of the Data Science application and its relevance to the job posting, however, it presented its own challenges such as the size of the data causing processing time to blow out quite dramatically when creating analyses on Jupyter Notebooks, as well as being unsuitable for Fairness Unaware Analysis.

2b

This feels like a quick task to get an MVP up and running for a more junior member of a Data Science team, or an experienced Data Scientist becoming familiar with an organisation and their data. Hypothetically speaking, out of the box models with default parameters could be suitable for production and achieving insights valuable enough for decisions that are able to be made within organisations. A minimal amount of effort has been exerted to get an idea of what types of models might be most effective, and some insights about the data have already been gained.

2c

Proposing a research topic or novel solution on this task is a little difficult as the data was labelled in spanish or had non descriptive, anonymised feature names. A research solution may be more focussed on the methodology, such as an exploration of the types of models that are used and the most beneficial accuracy measures that

could be used for data sets with disproportionate target classes, specifically in the domain of customer satisfaction. Customer satisfaction could also be further separated into additional classes and explored, for example customers who are satisfied and churn versus not churn, and customers who are unsatisfied and churn versus those who continue to use the service. Understanding these nuances could further the positive social impact of using particular Machine Learning in the context of customer service.