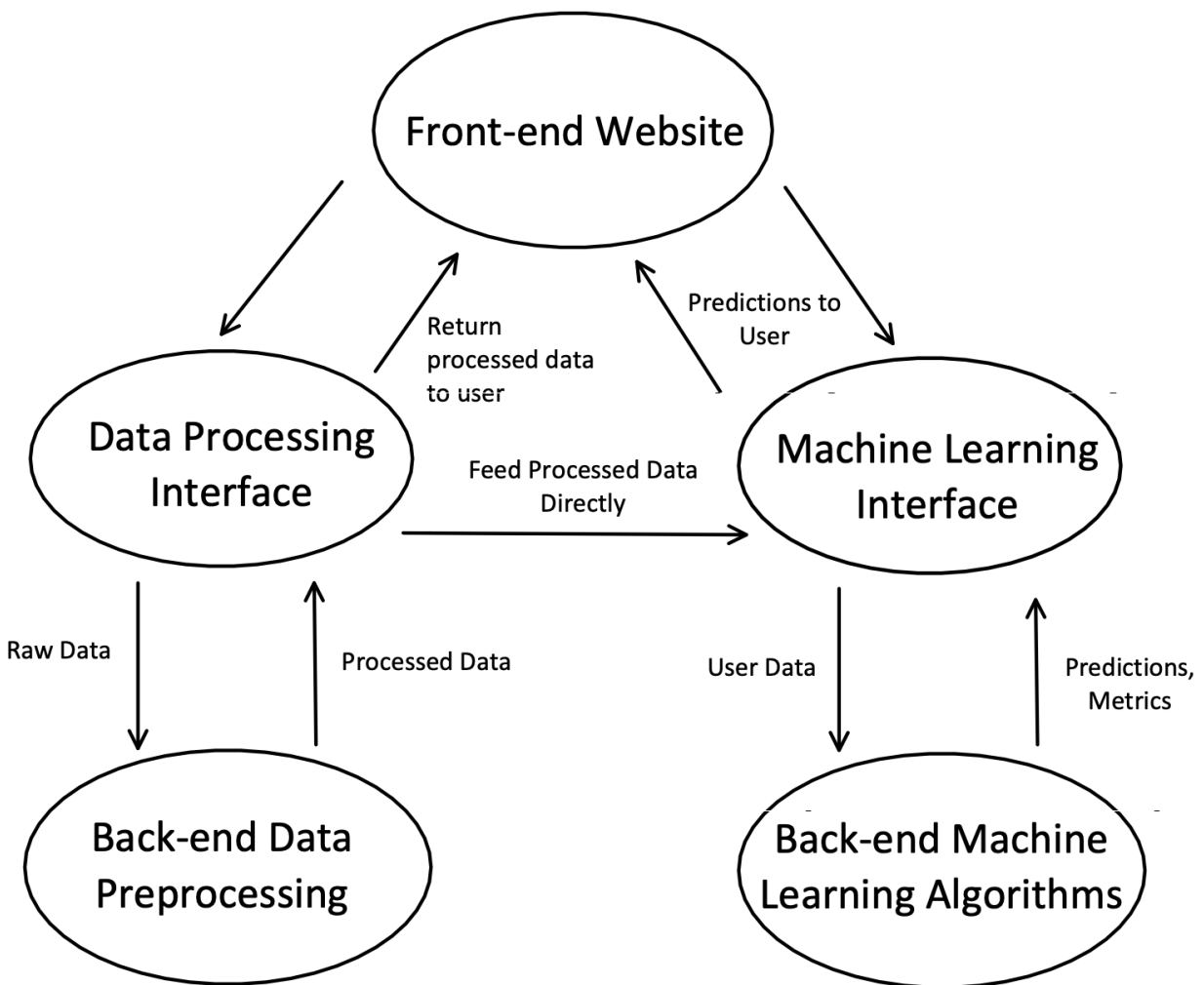# Dolphyn Small Business Insights

- Team Name
  - MathMajorsUnite
- Team Members
  - Georgia Channing
  - Harry Channing
  - Shannon Hall
  - Owen Queen

Introduction

Machine learning has become an increasingly prevalent tool in business from product recommendations to user feedback interpretation. Most of the organizations that moved quickly to integrate machine learning techniques into their workflows had the resources to hire machine learning engineers and data scientists to turn their data into actionable insights. Hiring machine learning engineers and data scientists is expensive and is out of reach for most small and medium businesses. That doesn't mean that they shouldn't be able to leverage machine learning to improve their businesses and make their lives easier. Enter Dolphyn: a one-stop shop for interpreting large scale customer feedback. Machine learning shouldn't be monopolized by Fortune500 companies, but should be available to small and medium businesses, who now more than ever need all the support they can get.

While there exist lots of plug-and-play machine learning libraries, Keras, sci-kit learn, TensorFlow, even the small amount of coding necessary to get models up and running is a barrier to entry. Our product will smooth the path to integrating machine learning into small business workflows by creating an easy-to-use interactive web platform. Our product will cut down on the overhead needed for the preprocessing steps of machine learning, and it will ideally give small businesses the tools they need to, in the future, incorporate machine learning into their workflow.

As we assembled our project, we realized that several pivots needed to be made: number processing should be done rather than image processing and focus should be on preprocessing and data understanding steps rather than on predictive steps. Our goal was to make a website that could both be informative, widely-used, and understandable. Focusing on image processing and prediction runs contradictory to our larger goals in that properly implementing either would both narrow the usability and interpretability of the website. This pivot had us largely focusing on the exploration of the existing data through standard data manipulation techniques; visualization of principal component analysis (PCA), t-stochastic neighbor embedding (T-SNE), and histograms; and Linear Regression.

Generally speaking, we completed our post-pivot project goals. We implemented data manipulation, PCA, T-SNE, Histograms, and Linear Regression in an INCREDIBLY easy to use format. If we were to move forward with the project, it would make most sense to first iterate on the explanatory information that is meant to be read in conjunction with the current output (to better understand what is going on). Then, we could further refine our product by creating more niche applications with different predictive models.

Customer Value

Our original "customer value" was that our product would be easy to use/understand by SMBs and help the said businesses implement machine learning tools in their workflow. As

addressed in the introduction, we later pivoted away from image processing (Sprint 3) and classification (Sprint 4) in favor of additional focus being put on preprocessing/data exploration. Our goal is to create a widely-applicable, easy to understand platform to guide the SMB workflow, and we felt that focusing on image processing and prediction would severely limit those attributes -- image processing is useful to only a small subset of businesses; different businesses need to implement very different model types. Focusing on preprocessing and data exploration would allow the site to both have more widespread use as well as inform customers about their data (instead of being unexplainable).

Technology

Our system can be thought of as front-end and back-end components. In the front-end component, our website will feature a polished GUI that allows users to easily process and analyze their data. In the back-end component, our system will perform preprocessing and analysis tasks. The front-end will serve as a friendly interface to the power of the back-end implementation, and a key portion of our project will be the integration of front-end and back-end components to achieve a seamless user experience.

The project was implemented primarily in Python. We will implement the front-end component in Django, a web framework implemented in Python. The back-end was implemented in various ML libraries including Scikit-Learn. To perform data preprocessing, we incorporated other classic data science libraries such as Pandas and Numpy. For database control, we used Django's built-in functionality as well as the Pandas and Numpy libraries to pass data between the front-end and back-end of the website. The website is only locally ran through Django's setup features; we decided to not try to implement it on Google Cloud or some other cloud-based platform due to time constraints. The design of the frontend was done mostly through Bootstrap, a free and open-source CSS framework.

Our website is easy to use and is quite effective at doing the things we want it to. We tested our interpretability of our site by having parents attempt to use it (with minimal guidance). It worked well in that environment. We tested the backend code by feeding in sample data sets from a variety of sources and performing different tasks on the said data-sets that would mimic use-cases customers of our product might have (PCA, T-SNE, Histograms, Linear Regression). While there are possibly some ways that users could break the site, we did not find any immediate vulnerabilities in the workflow of the site. Our site clearly does the jobs we set out for it to do. Here are some sample output from the tests:
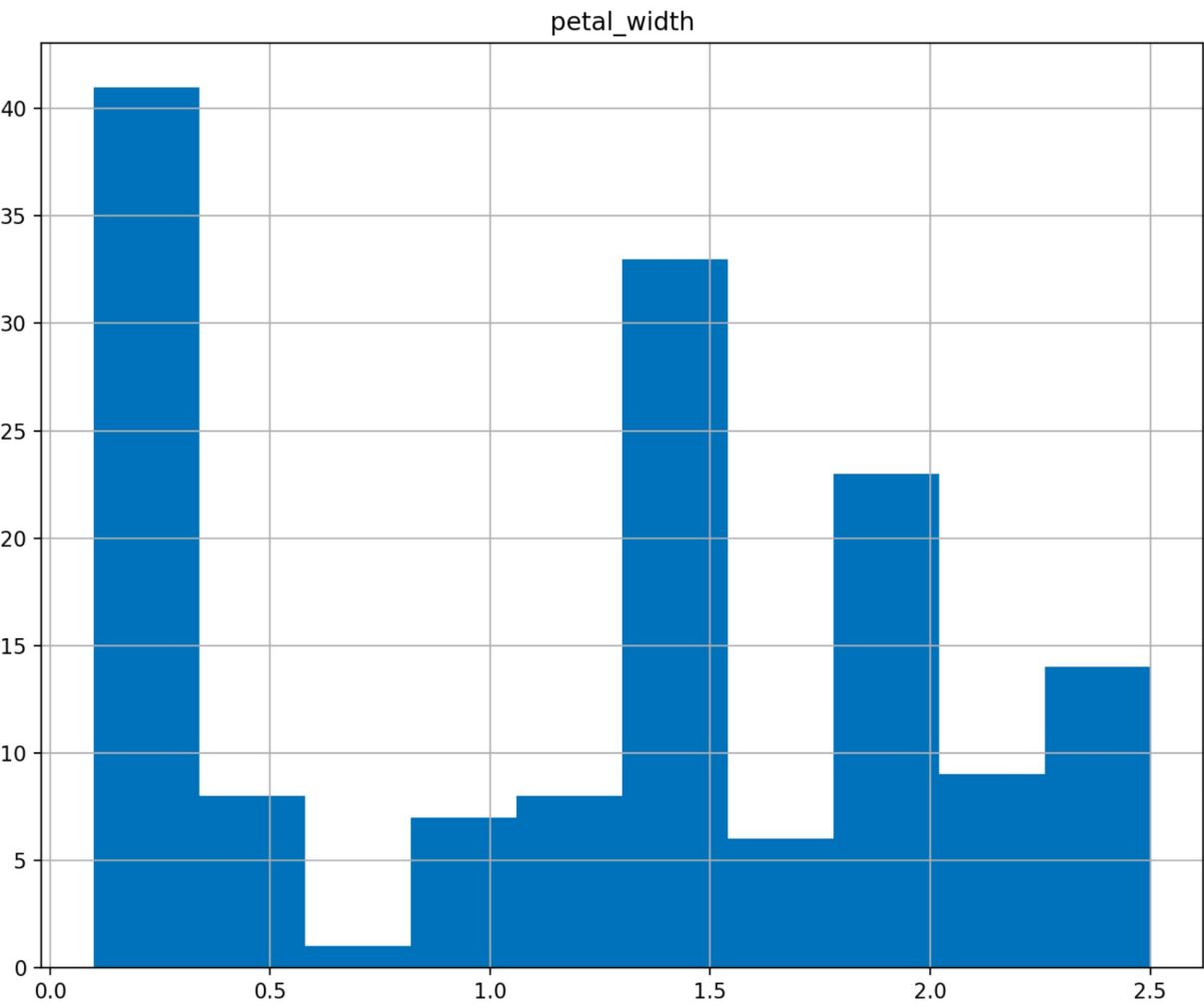
Predicting: petal_width

Variable	Coefficient
intercept	-0.01385
sepal_L	-0.08191
petal_length	0.44993

R^2 value = 0.929

Equation: petal_width = -0.01 + -0.08*sepal_L + 0.45*petal_length

Model generated has an R^2 of 0.93 on the given data!

petal_width

Team

Backend design was largely done by Ms. Channing with small contributions from Mr. Channing and Mr. Queen. Presentation development, graphical design, and communication between frontend/backend groups was largely done by Mr. Channing. Frontend UI was implemented by Mr. Hall. Frontend-backend merger was done by Mr. Queen. These roles were largely static throughout the project.

Going into the project, the roles and responsibilities were doled out so as to be equal (and siloed). This, unfortunately, became an unfair division of labor as it was much more difficult to learn how to and implement the frontend (Django) piece of the project than initially anticipated. As much of the coding was done late in the project, it also didn't make much sense for Mr. and Ms. Channing to attempt to learn Django as Django, as discussed earlier, has a significant learning curve. This resulted in Mr. Hall and Mr. Queen likely made the largest contributions to the project.

Full group meetings were organized by Mr. Channing and Mr. Queen. Frontend meetings were organized by Mr. Hall and Mr. Queen. We met as-needed throughout the entire semester, and most of our communication was done through texting and GitHub issues.

Project Management

While we largely completed all of the goals for our project, these goal completion dates were generally separate from our initial planned completion dates. Significant pivots were made which, of course, drastically altered plans. We also experienced some difficulties in the beginning-middle of the semester with different additional responsibilities (school, applications, and others) that made it difficult to effectively push code. This, of course, was followed by significant development happening in the last few weeks leading up to the project completion and presentation.

In addition, the learning curve for Django and HTML for front-end design of the website was quite difficult. Since no one in the team had prior experience with web development, it took longer for us to understand the object-oriented, jargon-laden design of Django, and this

undoubtedly delayed our progress on the project. Had we known how to use some of this technology before beginning the project, this project could have been much more extensive. However, we are pleased with our final product, and it accomplishes our basic goals and even more.

Reflection

There were many things that our group did well in maintaining throughout the project: group-dynamic, loyalty to business motivation, team communication, and setting attainable goals. At no point in the project were there large disagreements between team members, so there was no significant stoppage of work or scatterbrained code implementation. Another strength that was omnipresent within our team was the devotion to the initial product ideology (creating a usable, adaptable, explainable data-processing product). This devotion allowed us to pivot from our initial plan and make a more compelling product for our target customers. Team communication stayed clear and constant, allowing us to quickly address problems and shift plans as needed. Finally, we set attainable goals which allowed for more clear checkpoints and lowered overall stress levels. In tandem, the strengths of our team throughout the project is what lead to the completion and general success of our project.

There were a few difficulties that made the project more challenging than initially anticipated: siloed work efforts, busy schedules, and occasionally unclear next steps. At the start of our project, the doled out roles left Mr. Hall and Mr. Queen largely responsible for frontend development. Django proved significantly more difficult to pick up than say, the pickup of another object-oriented programming language. By the time they were in the thick of Django, it was unclear whether it would have been worthwhile for Mr. and Ms. Channing to move up the Django learning curve for the purposes of lending a hand. This resulted in additional time being put in by the said Django designers. Another difficulty in project implementation was that, in fact, all of us were quite busy for the meat of the semester with applications, research, internships, other classes, etc. so couldn't make as significant progress on the project as would have been optimal with a more diverse set of work schedules. Finally, at times, especially as we were pivoting, it was difficult to know what process should be implemented next. This resulted in occasional lack of direction and slowing in commits.

This project can largely be considered a success. In the business sense of creating an easily usable and widely applicable product for SMBs with nonexistent DS Departments, we followed through completely: we have a website that has been tested on ML-uninformed people (our parents) that has good and ~interpretable outputs (PCA, T-SNE, Histograms, Multivariate Linear Regressions). Going forward, if we were to continue the project, our next steps would be to further explain the meaning/output of our existing functions, then implement a variety of predictive tools for different use-cases.